ARI Research Note 2008-05

Training for Rapid Interpretation of Voluminous Multimodal Data

Dennis J. Folds, Carl T. Blunt, Raymond M. Stanley
Georgia Institute of Technology

Basic Research Unit
Paul A. Gade, Chief

April 2008

United States Army Research Institute
for the Behavioral and Social Sciences

Approved for public release; distribution is unlimited.

U.S. Army Research Institute for the Behavioral and Social Sciences

A Directorate of the Department of the Army Deputy Chief of Staff, G1

Authorized and approved for distribution:

Michelle Sams

MICHELLE SAMS, Ph.D.
Director

Research accomplished under contract for the Department of the Army

Georgia Institute of Technology

Technical review by:

Richard E. Christ, U.S. Army Research Institute

REPORT DOCUMENTATION PAGE

1. REPORT DATE (dd-mm-yy)	2. REPORT TYPE	3. DATES COVERED (from. . . to)
April 2008	Final	May 2002 - March 2005

4. TITLE AND SUBTITLE Training for Rapid Interpretation of Voluminous Multimodal Data	5a. CONTRACT OR GRANT NUMBER DASW01-02-K-0001 5b. PROGRAM ELEMENT NUMBER 611102A
6. AUTHOR(S) Dennis J. Folds (Georgia Institute of Technology) Carl T. Blunt (Georgia Institute of Technology) Raymond M. Stanley (Georgia Institute of Technology)	5c. PROJECT NUMBER B74F 5d. TASK NUMBER 2902 5e. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Georgia Institute of Technology 400 Tenth Street Atlanta, GA 30332	8. PERFORMING ORGANIZATION REPORT NUMBER A6849-07-001
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) U. S. Army Research Institute for the Behavioral and Social Sciences 2511 Jefferson Davis Highway Arlington, VA 22202-3926	10. MONITOR ACRONYM ARI 11. MONITOR REPORT NUMBER Research Note 2008-05

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
Contractor Officer's Representative and Subject Matter POC: Dr. Paul A. Gade

14. ABSTRACT *(Maximum 200 words)*:

Previous research has not specifically addressed rapid decision-making based on large amounts of data, although a large body of research has identified various biases and characteristic errors in human decision making that promote economy of information processing. The purpose of this series of experiments was to determine whether seven known characteristic error types operate in rapid decision-making, and to determine whether training to identify key contexts in which these errors are likely to occur can reduce their occurrence. Student volunteers across four experiments performed a simulated incident detection task using information in a variety of formats. Results supported the primary hypothesis that each of seven types of characteristic errors of interest occur in rapid decision making. A reduced tendency to commit false alarms occurs as a result of general alerting to the presence of error traps, although individuals who received specialized training to reduce the occurrence of characteristic errors performed no better than a control group. Team task performance greatly reduced undesirable false alarm errors, but also reduced desirable incident hits. Teams that received training reported the fewest hits and errors, although it was unclear whether this was a result of training or the effect of alerting observed in individuals.

15. SUBJECT TERMS
Training, decision-making, training evaluation, military training, decision biases

SECURITY CLASSIFICATION OF			19. LIMITATION OF ABSTRACT	20. NUMBER OF PAGES	21. RESPONSIBLE PERSON
16. REPORT	17. ABSTRACT	18. THIS PAGE			
Unclassified	Unclassified	Unclassified	Unlimited	64	EllenKinzer Technical Publicaiton Specialist 703/602-8047

ACKNOWLEDGMENTS

The present research was sponsored by the U.S. Army Research Institute (ARI) under contract DASW01-02-K-0001. The work was performed by researchers in the Human Systems Integration Division of the Georgia Tech Research Institute (GTRI) during the period May 2002 through March 2005. Dr. Dennis J. Folds was the GTRI Principal Investigator/Project Director for this effort. Dr. Paul Gade was the ARI Contracting Officer's Representative. Dr. Scott Graham, head of the ARI Field Unit at Ft. Benning, GA also provided significant technical input for this work.

Significant technical support during the contract period was provided by the following graduate students pursuing advanced degrees at Georgia Tech: Ms. Aideen Stronge, Mr. Richard Gordon, Mr. Kenneth Hailston, Ms. Kori Bevis, Ms. Catherine Jamal, Mr. Ray Stanley, and Ms. Courtney Lessl.

Excellent software support was provided by undergraduate coop students. The contributions of Mr. Ben Johnson are particularly noteworthy. This report documents the results of the four major experiments completed for this project.

TRAINING FOR RAPID INTERPRETATION OF VOLUMINOUS MULTIMODAL DATA

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Introduction

Military decision making in recent decades was characterized by staff-intensive, centralized procedures that depended on a serial flow of information from the field, through the commander and his staff, and back out to the field. In contrast, military decision making in the future is expected to be much more fluid, with greater emphasis on ongoing collaboration between the commander and personnel in the field. These changes in decision making are necessary to execute the adaptive and agile operations required for military effectiveness on the battlefields of the 21st century.

The changes envisioned in Army command and control in the next one to two decades will place far greater emphasis on rapid and flexible decision making by the on-the-ground commander of an Army Unit of Action. These decisions will be driven by information extracted from multiple data sources. These data sources will be a mixture of real-time sensors, recent intelligence products, archived data, voice communications, and of course the direct perception of the environment by the Soldier on the ground. In many cases there will be large amounts of data in different modalities and formats, and a major challenge for the Unit of Action will be to correctly determine the meaning and relevance of the data in a timely manner. The situation will be complicated by the fact that the data may be incomplete, inaccurate, ambiguous, and contradictory.

The incoming stream of voluminous data will be available to the on-the-ground commander and to others in the Unit of Action. Although the decision will be the commander's to make, the decision making process will be a function of a small group. Different individuals in the Unit of Action will have different roles to play in the decision making process. Some may have expertise in a specific technical area that is relevant to the current situation; others may simply be assigned to monitor data from certain information sources (such as remote surveillance assets).

The types of missions performed in the recent past are expected to dominate U.S. military operations in the future. These missions include peacekeeping operations, low-intensity conflict, asymmetric warfare in urban areas, anti-terrorism actions, and provision of humanitarian aid in areas of conflict. Traditional military missions that focus on capturing territory and destroying enemy assets will continue, but are expected to be far more rapid and limited in scope than in the past. Thus, the types of decisions made in future scenarios by the Soldier on the ground go far beyond maneuvering, targeting, and self-defense. Decisions will include detecting sabotage, determining whether to intervene in civil disturbances, identifying war criminals, and reacting to threats of terrorist acts. Many decisions will have to be made in real-time or near-real-time (i.e., immediately or very soon after receipt of data). In making these various decisions, the Army Unit of Action often will have access to large volumes of data—and will have to interpret, determine relevance, and assign meaning to that data in a timely manner, and use that information to make a response decision.

Research on human decision making in military contexts has not specifically addressed the problem of rapid decision making based on large volumes of data. Research on rapid military decision making has generally been focused on decision making under stress with incomplete and conflicting information, but has not emphasized large volumes of data in these situations. On the other hand, research that has examined interpretation of large data sets in military contexts has not emphasized real-time decision making. There is a need, then, for research that examines these two aspects of decision making jointly. Topics that need to be better understood include:

- Effects of data modality and format, with focus on the formats expected to be available to Army ground forces,
- Effects of the total number of sources and overall volume of data,
- Effects of the amount of relevant information (information density) and the extent to which the relevant information is incomplete, ambiguous, and inaccurate,
- The types of errors that individuals and small groups tend to make when dealing with these large data sets (and especially any systematic errors, or biases, that are characteristic of decision making in this environment), and
- Training methods that will help improve performance in this type of decision making.

Relevant Theoretical Perspectives

Initial attempts to describe human decision making in complex situations were based on formal or logical analyses, often with a strong mathematical basis. In general these attempts sought to characterize decisions in terms of *utility*. It was thought that such diverse activities as gaming wagers, national economic decisions, international political decisions, and military tactical decisions could be characterized in terms of the expected utilities in the cells of a payoff matrix. Although these attempts had some value in analyzing certain problems, they did not adequately describe actual decisions made in real-world (as opposed to laboratory) conditions.

One of the first successful opposing views was Simon's theory of satisficing (Simon, 1955). Simon believed that decision makers process options until a satisfactory option is found. This differs from utility analysis in that the optimal solution is not sought. Instead, the decision maker evaluates options until one is found that will meet some pre-defined criteria. That option is then implemented and the search for further options is halted. Although Simon's initial interest was in economic decisions, his perspective proved useful in analyzing decisions made in other situations as well.

The Recognition-Primed Decision (RPD) model developed by Klein (1989) built upon and significantly enhanced Simon's theory of satisficing. Klein conducted extensive studies of decision strategies in naturalistic settings. Contrary to most laboratory research on decision making, Klein found that the decision makers did not seem to consciously evaluate and weigh the advantages and disadvantages of available options. It was found that decisions made in the field were often subject to time constraints that did not allow the possibility of conducting a formal utility analysis. Instead participants often "relied on their abilities to recognize and appropriately classify a situation" in order to come to a decision (Klein, 1989).

According to the RPD model, the first question to be addressed in the decision process is whether the situation is familiar. If not familiar, the model predicts that decision makers will tend to reassess the situation and attempt to gather more information. If the situation is deemed familiar, then decision makers will recall the goals, expectancies, normal cues, and typical actions that are related to the situation. It is in this stage of the process that decision makers will tend to test their expectancies to determine if they have been violated. If the expectancies are violated, then decision makers will tend to reassess the situation. If the expectancies are not violated, then RPD predicts that the first typical action recalled from memory will be evaluated. If this action is determined to be viable, the next step in the process is implementation of the action. If the action is determined to not be viable, then the decision maker evaluates the next typical action.

According to Klein (1989), the RPD model has three predominate features. First, the model assumes that decision makers are capable of recognizing situations as being typical and, as such, each situation will have a set of typical response patterns. Second, the model assumes some degree of situational understanding. The decision maker must be able to understand the current situation in a broad sense. Third, the model assumes a serial evaluation of the recalled responses. Once a response is found that appears appropriate, the evaluation is terminated (as in Simon's theory of satisficing).

In order to develop recognition of a situation, Klein (1989) asserts that decision makers are required to have four types of information. First, decision makers must have a set of plausible goals before going into the situation. This suggests that briefing decision makers on the goals of the exercise is critical. Second, decision makers must have "critical cues" and "causal factors" available to them before the decision point (Klein, 1989). Both "critical cues" and "causal factors" should be identified and incorporated into training. Third, the decision maker should have a generalized set of expectancies. In other words, decision makers must have some expectancy of what impact their decisions will have. Finally, decision makers must be aware of typical actions relevant to the specific decision point. The decision maker must be aware of the possibilities that they will be asked to evaluate.

Relevant Research Findings

Research in team decision making goes back at least to studies performed in the 1960s by Fleishman and his colleagues (e.g., Fleishman & Harris, 1962). Much of the past research on team decision making has focused on team composition issues, such as the effects of gender and military rank of team members. The types of decisions that have been studied include jury deliberations and financial investment decisions. Some studies have used pseudo-military tasks, but in general there has been little direct study of military team decision making in a controlled research environment. This research has been useful in developing a better understanding of decision making processes within the small group, but does not address rapid decision making with large volumes of data.

The focus of research in the 1990s has been on tactical decision making under stress, in the aftermath of the *Vincennes* incident in which an Iranian commercial airliner was shot down. This line of research has been performed by Salas and his colleagues at the Naval Air Warfare Center/Training Systems Division (NAWC/TSD) and in contracts sponsored by that agency (see Cannon-Bowers & Salas, 1998). This program of research has generated many useful findings, especially with respect to training techniques and strategies. These studies have focused on rapid decision making in tactical situations, but have not explicitly addressed training for processing large volumes of data being rapidly received.

At Georgia Tech Research Institute (GTRI), we conducted a series of experiments on human decision making with large sets of multimodal data, in the context of civilian incident detection and management (e.g., see Folds, Fain, Beers, Stocks, Coon, & Ray, 1996; Folds, Fain, Stocks, Beers, & Ray, 1998; and Folds, Mitta, Fain, Beers, & Stocks, 1995). In these studies, participants processed data from multiple sources presented in differing formats to determine the presence and location of incidents, and to diagnose incident characteristics. In these studies we focused on developing operator support systems and on the division of responsibilities in a two-person team. We did not explicitly address training techniques for these tasks.

Biases in Human Decision Making

One promising set of findings that has emerged from several decades of research on decision making is the identification of various biases and characteristic errors in human decision making. Much of this work has been performed in the context of large data sets—particularly intelligence data. A general perspective on bias and error in decision making with intelligence data is presented by Heuer (1999). From his point of view, reasoning is primarily a result of unconscious processes that involve the detection and integration of sensory information, the recall of past experiences, and pattern matching. These unconscious processes are subject to biases that normally serve us well by allowing us to quickly come to conclusions about situations that might otherwise result in indecision or not coming to a decision quickly enough to carry out the appropriate action. Often "what appears spontaneously in consciousness is the *result* of thinking, not the *process* of thinking," (Heuer, 1999, emphasis added). The biases lie in the underlying processes, and are thus not immediately accessible to conscious monitoring through introspection.

These underlying processes of information processing may be viewed as an evolutionary adaptation to the large quantity of information gathered through the human sensory systems. To deal with this volume of information, humans developed a system of simplifying strategies that promote economy of information processing. The evolutionary advantage is that it allowed quick decisions to be made, perhaps avoiding danger or providing access to desired resources in a timely fashion.

In some cases, these unconscious strategies produce biases that result in errors in decision making. A prominent bias is that information that is concrete, personal, or vivid may be given more emphasis than information that is more abstract or less interesting (Nisbett & Ross, 1980). In other words, information that is collected first-hand is often given more credence than information derived from other sources. Similarly, information gathered from a description of an actual event is deemed to be more important than information derived from an analysis of trends or statistical data. Seeing is not always believing—especially when dealing with ambiguous stimuli. In addition, anecdotal evidence, while important, should not outweigh a mountain of historical data simply because the data are more abstract.

Another source of bias is the consistent failure to consider alternatives whose supporting evidence might not be directly available. For example, Fischhoff, Slovic, and Lichtenstein (1977) asked expert and non-expert mechanics to evaluate a performance failure using a complete or incomplete fault tree. Experts were less likely to correctly diagnose the fault if part of the fault tree was missing. Non-experts performed worse than experts. The authors concluded that both experts and non-experts failed to consider alternatives that were not directly in front of them and did not recognize that other alternatives existed.

Individuals tend to overemphasize the importance of the consistency of information. In many circumstances, the consistency of information is a good indication of its reliability if the information is derived from a variety of sources with many different perspectives. However, decision makers tend to consider the source of the consistent information in the development of their conclusions. This phenomenon has been called the "law of small numbers" by Tversky and Kahneman (1974). Basically, the authors conclude that individuals are not skilled in assessing the validity of information based on the sample size from which it was derived.

Gettys, Kelly, and Peterson (1973) describe a consistent failure of judgment in dealing with inconsistent or uncertain information. For example, when dealing with a report that may or may not be reliable, individuals tend to make a simple yes or no decision about whether to consider the evidence contained in the report (Heuer, 1999). Information that is deemed unreliable is wholly discredited and is not subjected to further scrutiny. Likewise, if individuals decide to accept the report, they tend to overestimate their confidence in the information.

Information that has been accepted has a tendency to influence decision making even if the information is later shown to be incorrect. In an experiment, participants were given fictitious information about their performance on an assessment task. Lau, Lepper, and Ross (1976) observed that the effect of the fictitious information persisted even after participants were told the original feedback was incorrect.

There also is a bias in favor of causal explanations of phenomena. In fact, human observers are so adept in discovering causal relationships, in the absence of any real explanation, they will often invent an explanation. The result is that randomness and chance are not adequately considered. Given a random distribution, human observers will tend to focus on a subset of the distribution that has a pattern. Often a conclusion will be drawn from this subset of data while ignoring the rest of the distribution (Kahneman & Tversky, 1972). Related to the bias of causal explanations is a bias toward perceiving events as being directed or organized by a centralized actor. Not all seemingly related events are the result of a coordinated effort.

Fischer (1970) noted the bias of assuming that the cause of an event is somehow related to its effect, and called this bias the "fallacy of identity." When dealing with physical properties, the idea of a relationship between a cause and its effect rings true. For example, a heavy object dropped from a significant height will tend to cause a large impact on the surface below. Heuer (1999) states that analysts tend to "reason in the same way under circumstances when this inference in not valid." For example, an observer might incorrectly assume that a financial crisis is the result of a financial event or that a large event must be caused by an equivalently significant action (Heuer, 1999).

The availability or retrievability of information, either from recall or the ability to imagine, often biases decision making (Tversky & Kahneman, 1973). For example, the likelihood that a particular hypothesis will be generated will be influenced by the recent generation of a similar hypothesis or a personal involvement in a comparable situation.

Finally, human observers tend to be heavily invested in early hypotheses. The individual tends to make adjustments to the original hypothesis rather than reconsider all of the evidence when new information is presented. This often results in the final judgment being closer to the original hypothesis than would be warranted by a second evaluation of the data.

Overall, the following biases in human decision making have been identified in research:

- *Confirmation Seeking Bias* – People try to confirm what they already believe even when a more credible explanation exists. They pay more attention to evidence that supports their view than evidence that contradicts their view.
- *Vividness Bias* – Information that people perceive directly (hear with their own ears or see with their own eyes) is likely to have greater impact than information received secondhand that may have greater evidential value.
- *Absence of Evidence Bias* – The absence of relevant information is not properly considered in decision making.
- *Availability Bias* – Decisions are often influenced by recent events or well-known conjectures; ability to imagine explanations not previously considered is limited.
- *Oversensitivity to Consistency Bias*–A lot of data pointing in one direction does not necessarily mean that the obvious conclusion is the right one. The information could have all been taken from the same source or subject to the same type of error. Multiple reports that in fact are derived from a single source may be treated as though they are independent confirmations of the observation.
- *Persistence of Discredited Information Bias* – Information that was deemed relevant often persists even after it has later been discredited.
- *Randomness Bias* – In general, there is a bias against defining something as random. Often people will impose a causal relationship where none really exists.
- *Bias of Centralized Direction* – People tend to perceive a centralized cause for related events.
- *Acquiescence Bias* – Once a critical mass of evidence has been obtained, further information tends to be ignored.
- *Segregation Bias* – The problem may be separated from its larger context to reduce complexity.
- *Sample Size Bias* – Evidence from small sample sizes is given equal weighting to evidence from larger sample sizes.

Each of the biases listed above can be interpreted as an unconscious strategy that serves to reduce the amount of information processing required to make a decision. It is not known whether these biases will become more pronounced when the decision maker is dealing with large volumes of data being rapidly received. The biases may be manifested in different ways, or new biases may emerge.

Training to overcome a bias is challenging because the biases are produced by unconscious strategies that are in some sense "natural" and in many cases are appropriate. Simply being aware that the bias exists is usually not sufficient to counteract the effects of the bias. One promising possibility for improving performance in decision making through training is to train individuals to spot the conditions that tend to produce the bias (i.e., characteristics of the stimulus environment) rather than attempting to eliminate the bias. This approach builds upon the human information processing strategies that produce the bias in the first place—using simplified "rules of thumb" to help parse data. The targeted response of the training will be learning to recognize the markers of biased decision making. Rather than trying to simply eliminate the bias in decision making, the decision maker could be taught to recognize the cues or markers of situations that produce the bias, and to take appropriate countermeasures against the bias.

For example, in examining a decision, the individual could be trained to recognize the need to use qualifying statements, such as "it all depends on…" or "it might be because of …"in defending their judgments. Such qualifiers often equate to a product of bias rather than careful analysis (Heuer, 1999). Individuals who minimize the use of qualifying statements and are able to fill in the details of the qualifier when used may tend to be more successful at overcoming the biases and rendering accurate judgments.

In summary, the research on the systematic errors that affect judgment suggests that decision makers should develop a method of detecting potential biases and take steps to counter the bias. They could be trained to weigh evidence in ways that will counter the bias. The individual also should develop a method of grouping related information. In addition, the source of the information should be maintained along with the information itself.

Research Questions

The overall goal of the proposed research is to increase understanding of how individuals assign meaning and relevance to large amounts of ambiguous data being rapidly received, and determine how to improve this ability through training. The specific objectives of the proposed research program are as follows:

- Assess the effects of data format, density, and overall volume on determination of relevance and assignment of meaning;
- Determine the biases and characteristic errors in determination of relevance and assignment of meaning; and
- Evaluate the effectiveness of anti-bias training as a method of improving performance in determination of relevance and assignment of meaning.

The specific objective for each major experiment is as follows:

- Experiment 1: Assess the effects of data format, density, and overall volume on assignment of meaning and on the types of errors and biases observed in these tasks, when performed by individuals.
- Experiment 2: Assess effectiveness of anti-bias training for individuals.
- Experiment 3: Assess the effects of data format on assignment of meaning and on the types of errors and biases observed in these tasks, when performed by small teams.
- Experiment 4: Assess effectiveness of anti-bias training for small teams.

Our primary hypothesis is that bias and characteristic errors observed in other types of decision making will be operative in tasks in which large volumes of ambiguous data must be processed rapidly, and that—if anything—the biases will become more pronounced in these tasks. The rationale is that the biases arise from unconscious strategies of economy in human information processing, due to the high demands of paying close attention to all streams of sensory data.

Our second hypothesis is that targeted anti-bias training can help overcome typical biases in decision making, thereby improving cognitive performance in situations where large volumes of ambiguous data must be processed rapidly. The mechanism that will be exploited for anti-bias training is the very mechanism that produces the bias in the first place, namely, recognition-primed decision making. The targeted response will be learning to recognize the markers of biased decision making. Rather than eliminate the bias, observers will be trained to recognize the potential for the bias and to take steps to challenge it.

Research Approach

A series of four experiments was conducted. The first two experiments were laboratory studies that examined individual performance. Experiments 3 and 4 were conducted with small teams of three, again in a laboratory setting. This section explains procedures and characteristics of the research common across these experiments. Anti-bias training also is discussed, although only participants in Experiments 2 and 4 received such training. A separate methods section is reserved in the discussion of individual experiments for explaining study-specific details.

Participants

Participants in Experiments 1, 2, 3, and 4 were largely volunteers recruited from the student body at Georgia Tech. Recruitment efforts were focused on ROTC students, with additional participants recruited from the general student population as needed. Informed consent was obtained from all participants, and participants were free to discontinue participation at any time without penalty. Students could choose to receive course credit as part of a Georgia Tech psychology course or $10 per hour payment as compensation for their time. Human participants research for this project, as for all research at Georgia Tech that involves human participants, was under the supervision of the Institutional Review Board (IRB) at Georgia Tech.

Apparatus

The experiment was conducted in the C4ISR simulation laboratory (C4ISIM) at Georgia Tech. Participants used a device called the Auxiliary Information Display (AID), driven by a background simulation running in the C4ISIM. The AID was an emulation of a generic hand-held (PDA-style) information appliance envisioned for use by Army infantry in the 2010 time frame. The AID served as a hand-held map, email device, photograph viewer, video player, and audio player. The AID was actually presented on a standard PC display, but was of the approximate size of a hand-held device. The C4ISIM provided information in a number of formats to the AID, enabling the participant to view (or hear) all of these information sources using the AID.

Figure 1 shows the AID in a map-view mode. The map is scalable and contained icons of selected activities or objects.

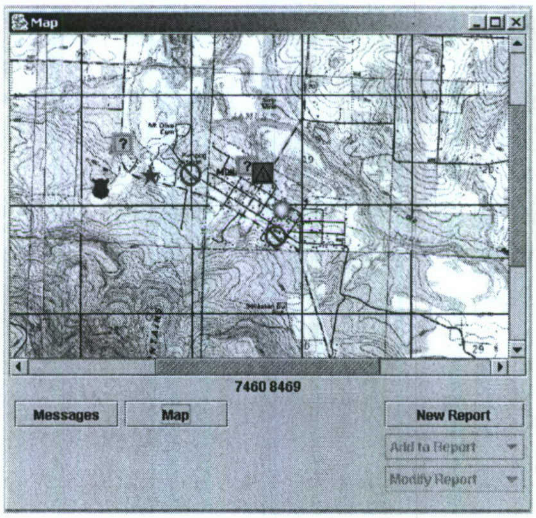

Figure 1. AID in map-view mode.

Figure 2 shows the AID in a folder-view mode. The basic scheme was an expandable/contractible folder list with an inbox for messages received, and archive folder for messages believed to be irrelevant, and additional folders for candidate incident reports created by the participant. The participant could access messages from the inbox, and move them to other folders as desired.

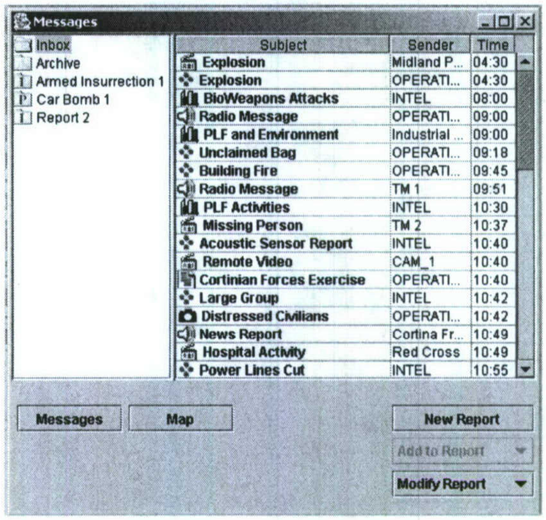

Figure 2. AID in folder-view mode.

The AID also had modes for viewing each type of information source (video playback, photo viewer, text viewer, and graph viewer). Audio playback was controlled from the folder view mode.

Procedure

The participant's primary task was to process a large number of data items in rapid succession and to report on the presence or occurrence of one or more target conditions (incidents). Participants were trained on the definitions of sixteen incident types, shown in Table 1.

Table 1
Incidents by Category and Type

Incident Category	Incident Type
Terrorist attack	Sniper Suicide bomber Bioweapons attack Hostage situation Kidnapping Information attack
Presence of target of military worth	Enemy military forces Bioweapons lab Arms cache Suspicious shipment Drug cache/shipment
Credible threat of terror	Car bomb Mine incident Disruption of public facility Bomb threat Insurrection/protest

Each incident type was defined in terms of key indicators (one or more of which had to be present to justify a report submission), and confirming indicators (which serve to confirm incidents in the presence of a key indicator). For example, to report the incidence of a sniper attack, the participant had to decide whether sufficient evidence existed for at least one of the following three key indicators provided in training:

- A gunshot is verified and the target appears to be a high-ranking government official (e.g., mayor, governor, president) or a key military leader (e.g., friendly forces unit commander, host nation unit commander, defense ministry official).
- Evidence that the shot(s) came from a vantage point (such as upper story window or rooftop). Examples include weapons signature or sighting of a shooter with a rifle.
- Information from intelligence sources of a confirmed sniper attack.

Additionally, the submission of an incident report required identifying at least one additional confirming indicator. Thus, each incident "event" present in a trial was supported by at least two data items: one key and one confirming indicator. Participants were allowed more flexibility in deciding which data items might serve as confirming evidence for a key indicator. A confirming indicator could be any data item that the participant decided helped to confirm the key indicator. This was often a data item that matched either a second key indicator for the incident, or one of the examples of confirming indicators provided in training:

- A higher incident rate of sniper fire in that area within the past year or when compared to surrounding areas.
- Evidence of the firing of a high-powered rifle (160-170 dB).

- Evidence of deliberate, selective, and accurate fire (e.g., individuals targeted out of a crowd, certain groups selectively targeted).

For any given trial, the participant's goal was to process every data item received in their inbox on the AID (shown in Figure 2) before reaching the time limit. The time allotted to perform a given trial was determined by the information volume of the trial. For the first experiment, all data items were present in the inbox at the beginning of the trial. For subsequent experiments, data items were delivered to the inbox as the trial progressed, at intervals equal to the difference in stimulus times (e.g., a stimulus item occurring at time 9:05 would be received approximately five minutes after a stimulus item occurring at time 9:00). One reason for the delayed delivery of items to the inbox in the subsequent experiments was that it allowed better control over items associated with one bias (persistence of discredited information) that required the presentation of one information item (the one to be discredited) prior to the presentation of subsequent items that discredited the earlier one. In the first experiment, the item to be discredited appeared higher (earlier) in the inbox, but nothing prohibited the subject from opening the later (discrediting) information items before opening the one to be discredited.

Upon accessing any data item from the inbox, the participant could take one of four actions: (1) start a new report submission folder, which resulted in the data item being moved to that folder; (2) add the data item to an existing report submission folder; (3) archive the data item, indicating that it is irrelevant (archived data items could be retrieved later), or (4) leave the data item in the inbox for continued consideration while other items are accessed. When the participant decided to submit a report, they either created a report submission folder, or opened a report folder that already existed.

Reports were submitted from the report-view on the aid (see Figure 3). To submit a report, participants had to provide the data items that supported the incident event, the incident category and type, the location, a time estimate, and an estimate of their level of confidence. Event locations and times were available from one or more data items associated with the event. In addition to the required data items, participants also could provide comments on the report, and could specify a name for the report. By default, a report name was provided based on the incident type provided by the participant (e.g., Sniper 1, Car Bomb 2, etc.). Once the participant had decided that the appropriate data items were in the folder and specified the location and time of the event, he/she could press the "submit" button to submit the report.

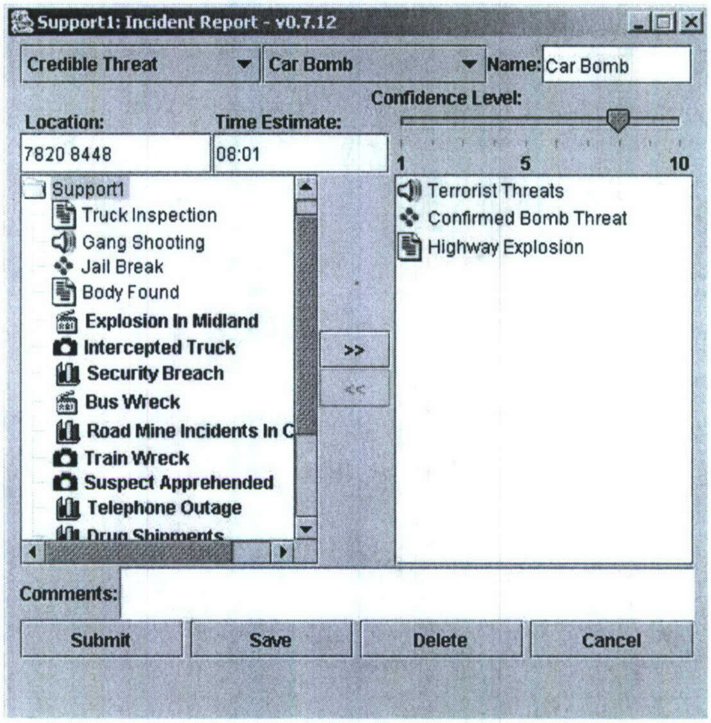

Figure 3. AID in report-view mode.

Stimuli

The stimulus set consisted of data items that varied by format (or modality) and type (incident, false alarm opportunity, or filler). There were six data formats used across experiments, examples of which appear in Figure 4. Audio items consisted of recorded voice communications that were no more than 20 seconds in duration. Icon items were displayed in the map view of the AID, showing the location of a specific entity or event. Clicking on an icon brought up a text annotation with a descriptor for what occurred, including the time and location. Image items were still photo images, usually with text stamp descriptor that included the time and location. Graphic items were computer-generated charts, graphs, or figures depicting data from non-visual sensors, or other forms of graphics showing relevant information. Text items were simple text messages that included such things as reports, orders, text of intercepted messages, and other intelligence data in text format. Finally, video items were recorded video clips of approximately 20 seconds in duration. As with the data items in the image format, video items often contained a text stamp descriptor that included the time and location of the scene. Although some image and video scenes were produced by the research team, the majority of these scenes depicted actual events. Participants were instructed to ignore the original context of a scene in the event that they recognized the event.

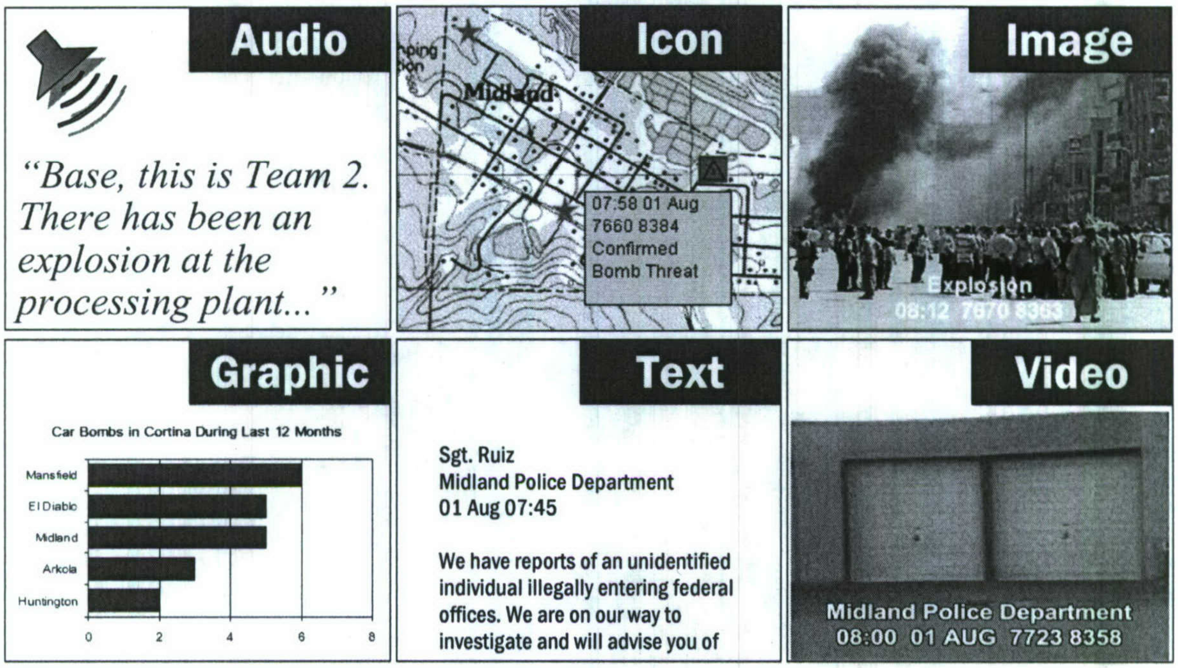

Figure 4. Examples of the six stimulus formats used across experiments.

Data items in the stimulus set were either associated with incidents or false alarm opportunities (FAOs), or were included in the set as Filler. As described above, incidents were the events participants were asked to report, consisting of at least two data items: one key indicator and one confirming indicator. All other data items fell into two general categories: distracters associated with FAOs (i.e., information that was designed to produce a false alarm if the participant did not correctly interpret the information) and filler distracters (information that was intended to add to the volume to the stimulus set, but that was not intended to support a bias or produce a false alarm). Filler items were designed to be similar in theme to items associated with incidents, and included such things as:

- Gunfire from law enforcement activities,
- Legitimate political gatherings and non-violent protests,
- Normal accidents and incidents in an urban area unrelated to enemy activity,
- Rhetorical threats of terror, and
- The presence of coalition forces and their military equipment.

FAO events were constructed to be similar to the characteristic patterns of information defined for incident events, but always lacked the key indicator item required for submission of a reportable incident. Thus, a FAO event related to a sniper incident might include data items for each of the examples of confirming evidence provided in training, but no key indicator. For FAO events, the pattern of information was configured to correspond to certain biases in decision making.

Seven of the eleven biases in human decision making discussed above were used to guide the creation of FAO events. The remaining four biases discussed (the acquiescence, bias of centralized direction, confirmation seeking, and segregation biases) were determined not to be amenable to study within the current context of rapid decision making. Each of the remaining biases was examined indirectly by identifying characteristic errors that might emerge in the process of performing the rapid decision making task. The following decision making errors are expected to occur if the associated bias is present in rapid decision making:

- *Vividness:* Interpretation of photographs and audio recordings (especially of speech communications) has a greater impact than is warranted.
- *Absence of Evidence:* The fact that certain information is not present, but logically should be present if some hypothesis is true, is not recognized.
- *Availability:* Information is deemed to be an instance of the most obvious explanation, although no direct evidence exists to support the conclusion.
- *Oversensitivity to Consistency:* Several pieces of identical information are interpreted as providing additional or corroborating evidence.
- *Persistence of Discredited Information:* Interpretation of an event persists even though a critical piece of evidence for the interpretation has been discredited.
- *Randomness:* Information is grouped to produce a likely sequence of events, although there is no proof that the information elements are related, or there is proof that the information elements are not related.
- *Sample Size:* Evidence from a small sample is given as much weight as (or is weighted higher than) contradictory evidence from a larger sample.

The stimuli were counterbalanced for modality of presentation, types of reportable incidents, and false alarm opportunities. Each trial presented one or more incidents, according to a somewhat-plausible storyline within a larger general scenario, and established conditions conducive to the manifestation of selected decision making biases. Each incident and FAO event was supported by two to six data items.

Variables

Independent variables of interest were overall volume of information (18, 30, or 60 data items in a trial), density of relevant information (3, 6, or 12 relevant data items in a trial), and false alarm opportunities (corresponding to the six characteristic error types identified in previous research.)

Reports submitted by participants were classified in three ways. An incident, or "hit," is a report of a true incident event that qualifies as one of the sixteen incident types provided in training. A false alarm opportunity (FAO) response is a report of an event experimentally constructed to resemble a true incident, but which contains insufficient evidence to warrant reporting. The final report type is the false alarm (FA), which is a report of an event that was not experimentally constructed and does not qualify as a true incident.

The primary dependent variables used in analyses are hit percent (the proportion of potential incident events that were submitted as reports) and FAO percent (the proportion of potential false alarm opportunity events that were submitted as reports).

In addition, an index of effective decision making pertaining to information use, known as the "report score," was computed as a measure of report submission performance. This metric is a reflection of whether messages are used accurately after being delivered, and can be thought of as a proportion of attachments that were relevant to the event. This metric was computed as follows:

$$\frac{\text{Number of event-related attachments}}{(\text{Number of event-related opportunities}) + (\text{Number of event-unrelated attachments})}$$

The number of event-related opportunities refers to the number of data items experimentally designed to be associated with an incident or FAO event. The number of event-related attachments refers to the number of event-related data items that were submitted with the report. The number of event-unrelated attachments refers to the number of data items that were submitted with the report that were not experimentally designed to be associated with the event. Values for this index range from 0 to 100, where a score of "0" indicates no report was submitted for the event, and a score of "100" indicates that a report was submitted with (and only with) all experimentally defined data items for the event.

As the number of event-related attachments increases, the report score increases; as the number of event-unrelated attachments increases, the score decreases. For reports of incident events, a higher score suggests better performance. For reports of FAO events, a higher score is suggestive of greater susceptibility to the associated decision bias. Theoretically, an extremely large number of extra attachments also could push the report score to zero.

Experiment 1

The purpose of Experiment 1 was to determine whether seven specific characteristic biases observed in other contexts operate in this rapid decision making task, and to determine whether other biases also emerge. The experiment was designed to assess the effects of data format, density, and overall volume on assignment of meaning and on the types of errors and biases observed in these tasks when performed by individuals.

Experiment 1 Method

Participants. Participants for Experiment 1 were 21 volunteers from Georgia Tech, the majority of which were ROTC students.

Procedure. Participants received two days of training followed by two days of testing. The training days consisted of a basic orientation to the task and to the data formats used in the task, and practice performing the task. No anti-bias training was provided in this experiment. The test day sessions each consisted of a two-hour block of nine trials. At the beginning of each trial, a set number of items, dependent upon the volume for that trial, were in the inbox, ready to be accessed.

The time allotted to perform a given trial was determined by the information volume of the trial, and was based on an average of 20 seconds per data item, plus one minute for reporting. Low volume trials featured 18 data items for which participants were allowed to take up to 7 minutes. Medium volume trials featured 30 data items for which participants were allowed to take up to 11 minutes. High volume trials featured 60 data items for which participants were allowed to take up to 21 minutes. Trials also varied by the density of reportable items. Low density trials had 3 relevant data items, medium density trials had 6 relevant data items, and high density trials had 12 relevant data items.

Stimuli. In addition to volume, trials varied by density of reportable items (3, 6, or 12), with 1, 2, or 4 incidents present in each trial (trials with 3 incidents present were used in training, but not in testing). The first block of nine trials was used for 1 day of testing, and the second block of nine trials was used for the other day; which block of trials was completed first was counterbalanced. Trials within each block were presented in a random order for each participant. The items for both volume and density were distributed evenly across data item modalities.

Experiment 1 Results

Event type, volume, and density on event reporting. One individual was excluded from the analysis of report score by volume and density due to technical problems resulting in missing data.

Scoring criteria for the two major event types (incidents and FAOs) were that the specified event location was within a given radius of the actual event location, and that the report be submitted before the end of the trial. A correct incident report (hit) was scored if the participant filed an incident report that corresponded to a reportable incident in the trial. Reports that incorrectly specified incident type or that attached incorrect supporting elements were still scored as a hit. The participant committed a false alarm if an incident report was filed that did not correspond to a reportable incident. A false alarm was scored if the incorrect incident report specified an incident within a defined radius of the FAO location. Virtually all false alarms observed in the experiment were associated with FAOs. Hit percent and FAO percent were calculated for each trial, as were report scores for both event types. The former represent the proportion of events for which an incident report was submitted, and the latter reflect the degree to which information in the report agrees with the data items assigned to that report. Reports filed with a subset of associated event items, or that included items not associated with the event, received a report score penalty. The average report score will always be less than or equal to the proportion of events submitted.

Overall, the proportion of incident events ($M = 0.77$, $SD = 0.11$) reported was more than twice that of FAO events ($M = 0.35$, $SD = 0.17$). Report scores for incident and FAO events were 0.67 ($SD = 0.11$) and 0.29 ($SD = 0.13$), respectively.

Hit percent and FAO percent for each level of trial volume and density are presented in Table 2. The top section summarizes results for incident events. No sizable differences in hit percent are evident across levels of density, but the hit percent on high volume trials is slightly higher than that observed on the low and intermediate volume trials. A 3x3 repeated measures ANOVA indicated an effect of volume $F(2, 38) = 5.706$, $p = .007$, but no effect of density $F(2, 38) = 0.970$, $p = .388$, or their interaction $F(4, 76) = 0.346$, $p =.846$, on hit percent.

Report scores for incident events are presented in the top section of Table 3. For the most part, report scores are close to the associated mean hit percent, indicating that participants were fairly accurate in what was reported (the correct data items were filed with reports that were submitted). Accuracy was not as high on the high volume trials, as the report score is lower when compared to hit percent, when compared with means for low and intermediate volume trials. A 3x3 repeated measures ANOVA indicated no effect of volume $F(2, 38) = 1.525$, $p = .231$, density $F(2, 38) = 0.241$, $p = .787$, or their interaction $F(4, 76) = 1.183$, $p = .325$, on hit percent.

Table 2

Mean Proportion (SD) of Incident and FAO Events Submitted, by Volume and Density

Condition	Volume 18	Volume 30	Volume 60	Total
Incident Events				
Density				
3	0.75 (.34)	0.68 (.29)	0.83 (.24)	0.75 (.19)
6	0.81 (.18)	0.73 (.24)	0.85 (.19)	0.80 (.13)
12	0.71 (.19)	0.72 (.19)	0.82 (.19)	0.75 (.13)
Total	0.76 (.18)	0.71 (.13)	0.83 (.15)	
FAO Events				
Density				
3	0.53 (.23)	0.26 (.24)	0.23 (.21)	0.34 (.17)
6	0.38 (.21)	0.36 (.18)	0.61 (.20)	0.45 (.17)
12	0.20 (.24)	0.30 (.23)	0.27 (.20)	0.26 (.19)
Total	0.37 (.18)	0.31 (.17)	0.37 (.17)	

Results for FAO events are summarized in the bottom section of Table 2. The FAO percent was highest in trials that featured intermediate density of relevant information and high overall volume—the FAO percent was about 60% across these trials. A 3x3 repeated measures ANOVA indicated a significant interaction $F(4, 76) = 17.704$, $p < .001$ and main effects of volume $F(2, 38) = 3.390$, $p = .044$ and density $F(2, 38) = 25.340$, $p < .001$, in reporting false alarms. Whereas the FAO percent is highest on intermediate density for volumes of 30 and 60, across the lowest level of volume FAO percent is highest on low density trials, decreasing with increasing density. The reason for this difference is not clear. Report scores for FAO events are presented in the bottom section of Table 3. The pattern of means is similar to those for FAO percent, with a 3x3 repeated measures ANOVA FAO report scores indicating a significant effect of density $F(2, 38) = 30.273$, $p < .001$, and the interaction between density and volume $F(4, 76) = 16.191$, $p < .001$, in reporting false alarms, but no effect of volume, as indicated by a 3x3 repeated measures ANOVA.

Table 3
Mean Report Score (SD) of Incident and FAO Events, by Volume and Density

Condition	Volume 18	Volume 30	Volume 60	Total
		Incident Events		
Density				
3	0.74 (.34)	0.62 (.26)	0.69 (.22)	0.69 (.17)
6	0.74 (.17)	0.65 (.22)	0.67 (.21)	0.69 (.14)
12	0.62 (.18)	0.65 (.19)	0.72 (.19)	0.67 (.13)
Total	0.70 (.18)	0.64 (.12)	0.69 (.15)	
		FAO Events		
Density				
3	0.40 (.19)	0.23 (.21)	0.15 (.16)	0.26 (.13)
6	0.33 (.17)	0.33 (.16)	0.55 (.19)	0.40 (.15)
12	0.19 (.22)	0.23 (.19)	0.21 (.15)	0.21 (.16)
Total	0.31 (.15)	0.26 (.15)	0.30 (.13)	

Decision-making biases. False alarm percent also was compared across bias type. Results are shown in Table 4. FAOs associated with the oversensitivity and vividness biases were more likely to be reported than FAOs associated with other biases, and FAOs associated with the sample size bias were least likely to be reported. A one-way repeated measures ANOVA indicated an overall effect of bias type $F(6, 120) = 7.503$, $p < .001$.

Table 4
False Alarm Opportunity Percent by Bias Type

Bias type	M	SD
Oversensitivity	0.50	0.21
Vividness	0.45	0.29
Availability	0.37	0.21
Absence of Evidence	0.34	0.21
Randomness	0.30	0.17
Sample Size	0.27	0.22
Persistence of Discredited Information	0.26	0.13

Item modality. The potential influence of data item modality on event reporting was examined by comparing the proportion of potential data items reported for each item modality. For true incident events, relevant text, video, and image files were most often included with incident reports (Figure 5). A one-way repeated measures ANOVA for incident data item inclusion indicated an overall effect of modality $F(2.231, 44.615) = 19.709$, $p < .001$ (degrees of freedom reflect Greenhouse-Geisser adjustment due to lack of sphericity in the data). For FAO events, image, graphic, and icon files were most often used in reports (Figure 6). A one-way repeated measures ANOVA also indicated an overall effect of modality $F(2.581, 51.614) = 20.961$, $p < .001$ (degrees of freedom reflect Greenhouse-Geisser adjustment due to lack of sphericity in the data). Audio items often were not included for both types of events.

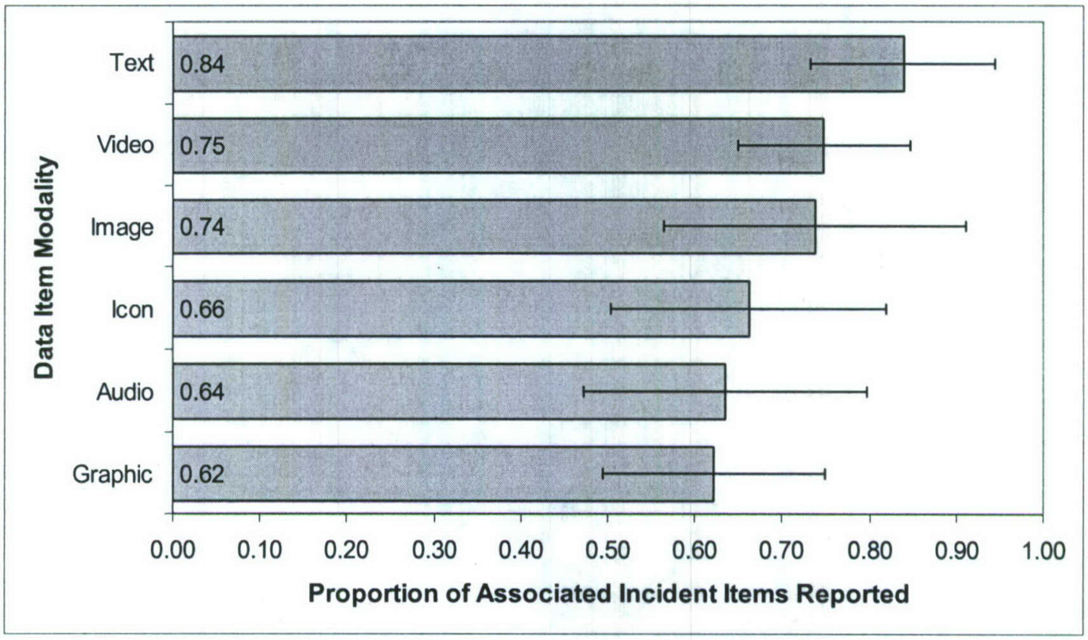

Figure 5. Proportion of Experiment 1 incident items submitted, by modality.

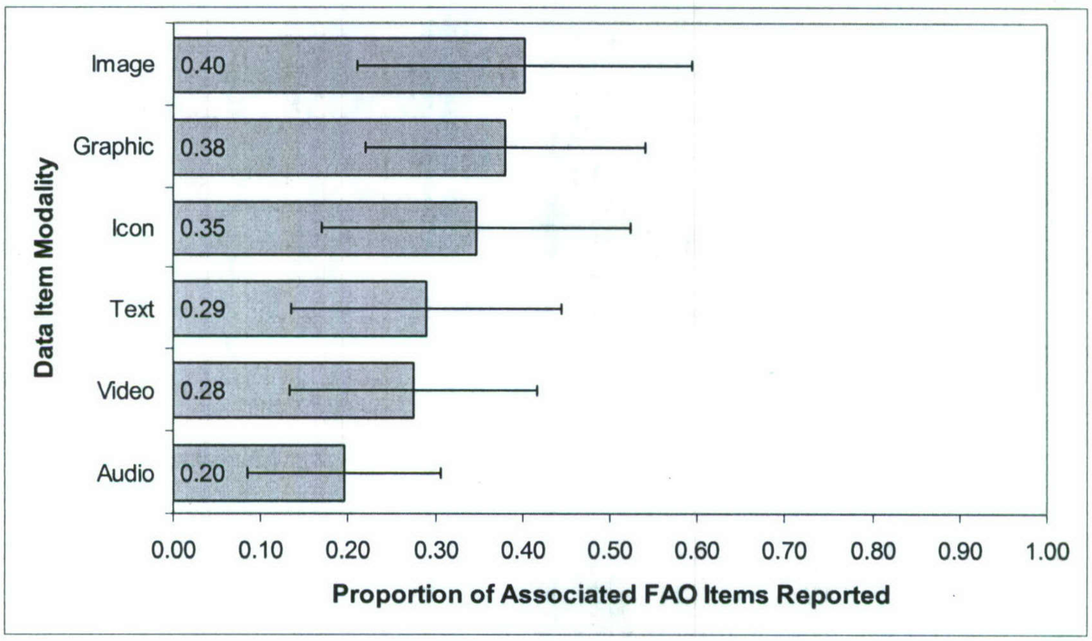

Figure 6. Proportion of Experiment 1 FAO items submitted, by modality.

Experiment 1 Discussion

Experiment overview. Experiment 1 was the initial investigation into the role that cognitive biases play in rapid decision making within a military context. It also considered the effect of density and volume on reporting, as well as which modalities were submitted most often with reports.

Event type, volume, and density. Overall, participants were about twice as likely to report incidents than false alarm opportunities. Thus, participants were generally able to discriminate the incidents, which they should report, from the false alarm opportunities, which they should not.

On incident events, an effect of volume indicated that the proportion of incident reports differed depending on the number of items processed in the trial, with the highest proportion of incident events reported on large volume trials. Performance did not differ across volume, however, when comparing report scores. This suggests that incidents were so clearly indicated that the span of volume and density values were not wide enough to reduce performance.

The density and volume values in this experiment did seem to influence performance on FAOs. The main effect of density shows that participants were more likely to commit errors associated with FAOs at the middle level of density than any other. This is probably due to the interaction, with an extremely high FAO percent at the highest level of volume and at the middle density level. The effect at the middle density on FAOs suggests that a certain optimal amount of total information (not too much, not too little) increases the likelihood of committing decision errors. However, this particular effect at the middle density is difficult to interpret and may be an anomaly.

One might expect the FAO percent to increase with decreasing density and increasing volume, because the increase in irrelevant information would increase the chance of making mistakes, and the data items to be less carefully considered. The opposite trend, however, was found in the present data. Considering the ambiguous nature of the data demonstrates a plausible explanation for why this may happen. As more information is given to process, there are fewer resources for a participant to make a final decision on whether a cluster of information is really an incident or not. Rather than submitting a false report, the reduced accuracy may be reflected by not submitting anything, even though the participant thinks something might be there; that is, a report folder may even be formed, but there is not enough time to evaluate the evidence and make a final decision to submit a report.

Decision biases. Of particular interest in Experiment 1 is the effect of bias type on FAO percent. The results suggest that all seven types of characteristic errors of interest in the present research occur in rapid decision making, although perhaps to varying degrees. For the present experiment, participants were especially susceptible to the oversensitivity and vividness biases. The first suggests that in a rapid decision making context, people are prone to judge information to be more credible simply because it is received from multiple sources. The second suggests that people are inclined to place greater evidential value on vivid sources of information (audio, video, image files) than is warranted by the content.

Item modality. For reports of both incident and FAO events, the tendency for participants to include an associated item differed based on the item modality. For incident reports, text files were most often included, followed by video and image files. For FAO reports, relevant image and graphic files were most likely to be included. Although the stimuli were counterbalanced with respect to modality, the complexity of the experimental design did not allow for counterbalancing with respect to such things as item importance and bias type. For example, the type of information supplied by graphic items was often the type required for confirming evidence, and only very rarely do graphic items serve as key indicators for an event. Further, certain FAO events require items of specific modalities making counterbalancing impossible. The vividness bias is based on audio, image, and video files, whereas the sample size bias is more likely to occur in the context of graphic, text, and audio sources of information. The design limitations make it difficult to interpret the importance of the reported results.

Of most interest is the difference in modality proportions observed across incident and FAO events. Where graphic items were the least likely to be included in incident events, they were very common in reports associated with FAOs. As mentioned earlier, graphic items serve well as confirming indicators. Graphic data can often be interpreted in many different ways, depending on what the user is looking for. This ambiguity may decrease its chance of being included in incident reports (especially if other confirming evidence is available), but increase the chance of decision making error based on the vividness bias. The decision makers viewed a vivid data item that made them want to make a report, and then saw confirmation of the event in the graph because they were looking for it.

Experiment 2

Experiment 1 confirmed that errors in decision making consistent with associated decision making biases are observed in the rapid decision making task. Experiment 2 evaluated the effect of training individuals about markers of the decision-making biases confirmed in Experiment 1. The experiment was designed to assess the effects of anti-bias training, data format, density, and overall volume on assignment of meaning and on the types of errors and biases observed in these tasks when performed by individuals.

Experiment 2 Method

Participants. Participants for Experiment 2 were 18 volunteers recruited from the student body at Georgia Tech, the majority of which were ROTC students. Most participants previously participated in Experiment 1. Participants who participated in this experiment but did not participate in Experiment 1 were given extended introductory training so that their experience with the task was approximately equivalent to the participants who continued from the first experiment. Participants were randomly assigned to either the training or practice group, each consisting of two, 2 hour blocks.

Training. In Experiment 2, half of the participants were trained to recognize conditions prone to decision-making biases, or "markers." Participants in the anti-bias training group received instruction on how to identify these markers in order to avoid making a decision error indicated by the associated bias. Anti-bias training began with the experimenter explaining that there were "information traps" that people are prone to when making decisions rapidly, which are due to "rules of thumb" that people form "to simplify mental processing." Participants were given binders with all the training information in writing so that they could follow along. Next, the experimenter explained to the participants that they would be receiving some training on these "information traps," and that in the experimental task they would be asked to verbally indicate when these "traps" arise. They were instructed to make these statements into a microphone, and were told that the experimenter would record their observations. They were instructed to specify which trap was occurring and say which pieces of information they felt created the trap.

The experimenter then explained each of the seven biases implemented in the experiment. For each bias, the experimenter provided an abstract description of the bias, a real-life example of the particular information trap, an example more relevant to the experimental task, and a description of indicators or markers of the information trap. Then, an abbreviated trial was run with the AID, and participants were asked to describe how the particular trap was formed. These practice trials were similar to real trials, with multiple events. After these steps were completed for each bias, participants were given a written quiz where they had to match the indicator of each with the name of each bias. After completion of this quiz, the experimenter reviewed the answers with the participant. Finally, two practice trials were completed where participants completed the task the way they would in the experimental task—submitting reports and verbally reporting the traps.

Each of the biases was given an appropriate "nickname" to facilitate learning of the bias. A sheet with the following summary information was available at all times during task trials:

- *Missing Link (Absence of Evidence):* The missing link trap occurs when the absence of relevant information is not properly considered in decision making; meaning, a critical piece of information you should have to report an incident with confidence is missing.
- *Familiarity (Availability):* This occurs when current events or personal experiences make it difficult for you to consider alternative explanations. People are predisposed to make a certain assumption based on recent events, based on the ability to retrieve similar instances from memory or to imagine an event could happen under various circumstances. In the circumstances of this experiment, if you allow current or recent events to influence your decisions, you may commit an error in reporting an incident.
- *Redundancy (Oversensitivity):* The redundancy trap occurs when the decision to file a report is made simply because there are several pieces of information that say the same thing (discuss the same aspect of an occurrence), when in fact the information could have been taken from the same source and therefore be subject to the same type of error.
- *Stubbornness (Persistence of Discredited Information):* The stubbornness trap occurs when a report is filed that includes information that is discredited (proven to be incorrect) or contradicted by other information. Believed accuracy often persists even after the beliefs have been discredited.
- *Coincidence (Randomness):* The coincidence trap occurs when a report is filed using completely unrelated pieces of information because there is believed to be a causal explanation, when in fact a common cause does not exist. This may result, for example, from simultaneous similar events that occur or an event occurring on the same date as a past significant event in a different year, when neither has causes relating them.

- *Small Sample (Sample Size):* The small sample trap occurs when a report is filed using evidence from small sample sizes that contradicts evidence from larger sample sizes. Sometimes small samples are treated as equally important as large samples, though the evidence they provide is not as strong as that from a large sample. Weighing both groups the same (attributing equal importance to both) can result in an error in judgment, particularly if the small sample is inadequate for drawing conclusions.
- *Seeing (or Hearing) is Believing (Vividness):* This occurs when a report is based solely on more vivid information (generally observed directly). Information that is vivid, concrete, and personal usually has a greater impact on decision making than abstract information. For example, because you can visualize an event better through video, audio, and/or image messages, these may influence your perception of the event more than a graph or text, even though they may be inconclusive in and of themselves. Similarly, if a report is received directly from an individual as opposed to one reporting on someone else's observations, the direct report may be a stronger influence in your reporting.

The participants who did not receive the anti-bias training underwent a similar procedure without specific information about the biases. The control group training began with the experimenter explaining according to a script that people make mistakes when rapidly making decisions, and that participants may notice "ambiguities or other situations that may present an opportunity for error." Participants were informed that their task would be to verbally indicate "error opportunities" and explain what about the stimuli made them come to that conclusion. The participants then practiced on seven abbreviated trials, the same trials used for the training group. For each trial, participants were to evaluate the messages and file a report, if appropriate; if not filing a report, they were to indicate why. If not identified, the experimenter highlighted how aspects of the information did not quite warrant report submission; these were the same explanations provided to the training group, without direct identification of the bias.

Procedure. The experiment was divided into four sessions. In the first session, participants were provided with a refresher on the task and interface. This was similar to the training they received as participants in Experiment 1. In addition, participants were alerted to the possibility of "error opportunities" and were trained on how to provide verbal reports to the experimenter when an opportunity for error was discovered. Participants in the experimental group received the anti-bias training described above, and were asked to provide verbal reports of any "information traps" they were able to identify, along with which pieces of information (data items) served as indicators (markers) of the trap. Participants in the control group did not receive anti-bias training, but were simply alerted to the possibility of "error opportunities" and were instructed to report "ambiguities or other situations that may present an opportunity for error." Thus, although both groups were alerted to the possibility of potential error opportunities, only the experimental group received training on how to identify markers for such opportunities. Both groups performed a series of seven short trials to practice providing verbal reports to the experimenter. As in Experiment 1, session 2 consisted of practice trials, and sessions 3 and 4 were test day sessions that each consisted of a 2-hour block of 9 trials.

Finally, Experiment 2 (and all subsequent experiments) differed from Experiment 1 in the delivery of individual data items. In Experiment 1, data items were all available from the beginning of the trial. In Experiment 2, data items were available on an experimenter-controlled schedule as the trial progressed. This allowed for a sporadic delivery of stimulus material that more closely simulated a real tactical command center. It also provided direct manipulation of the order in which information was accessed, which is important for the implementation of some of the biases that involved order of receiving information (i.e., persistence of discredited information).

Stimuli. The stimulus set for Experiment 2 was envisioned to be similar to that of Experiment 1, but because participants had previous experience in Experiment 1, a new set of stimuli had to be developed. Although the intent was to produce trials covering the same levels for the independent variables of data format, density, and volume, it was determined post hoc that some trials had inadvertently repeated location information for other events. Hence, there were five levels of density in Experiment 2 (3, 6, 9, 12, 14) rather than the planned three, and levels of density were not appropriately counterbalanced across trials. Further post-hoc analysis revealed errors in the construction of FAO events for Experiment 2, where the characteristic pattern for a FAO either pointed to more than one bias type (ambiguous) or none at all (spurious). The impact of this problem is discussed in the results.

Experiment 2 Results

Event type, volume, and density. Analyses for Experiment 2 parallel those for Experiment 1, but with the added variable of training condition. The overall hit percent was 0.63 (SD = 0.12), and was higher for the experimental group (M = 0.66, SD = 0.14) than the control group (M = 0.60, SD = 0.11), although the difference was not significant, $F(1, 17) = 0.960$, $p = .341$. The overall FAO percent was 0.14 (SD = 0.09), and was again higher for the experimental group (M = 0.16, SD = 0.10) than the control group (M = 0.12, SD = 0.09) but not significant, $F(1, 17) = 0.614$, $p = .444$. The average report score for incident events (M = 0.54, SD = 0.11) was higher than for FAO events (M = 0.10, SD = 0.07), $F(1, 17) = 610.38$, $p < .001$.

Hit percent and FAO percent are summarized by volume in Table 5, with report scores in Table 6. The top section summarizes results for incident events. No sizable differences in hit percent are evident across levels of volume, and no statistically significant effect of volume, training, or their interaction was revealed with a 2 (training condition) x 3 (volume) mixed ANOVA, $F(2, 34) = 3.101$, $p = .058$. The same was true for mean report scores of incident events across volume and training condition, $F(2, 34) = 2.760$, $p = .078$.

Table 5
Mean Proportion (SD) of Incident and FAO Events Submitted, by Volume

Condition	Training Condition: Control	Training Condition: Experimental	Total
Incident Events			
Volume			
18	0.59 (.13)	0.64 (.17)	0.62 (.15)
30	0.66 (.17)	0.70 (.12)	0.68 (.14)
60	0.55 (.16)	0.63 (.19)	0.59 (.18)
Total	0.60 (.11)	0.66 (.14)	
FAO Events			
Volume			
18	0.06 (.10)	0.07 (.07)	0.07 (.08)
30	0.11 (.10)	0.22 (.16)	0.17 (.14)
60	0.17 (.11)	0.16 (.09)	0.17 (.10)
Total	0.12 (.09)	0.16 (.10)	

Table 6
Mean Report Score (SD) of Incident and FAO Events, by Volume

Condition	Training Condition: Control	Training Condition: Experimental	Total
Incident Events			
Volume			
18	0.51 (.12)	0.58 (.14)	0.55 (.13)
30	0.55 (.14)	0.57 (.11)	0.56 (.12)
60	0.45 (.13)	0.54 (.16)	0.49 (.15)
Total	0.50 (.10)	0.56 (.11)	
FAO Events			
Volume			
18	0.04 (.07)	0.06 (.07)	0.05 (.07)
30	0.10 (.08)	0.18 (.14)	0.14 (.12)
60	0.11 (.07)	0.11 (.06)	0.11 (.06)
Total	0.08 (.06)	0.12 (.07)	

Data for FAO events across volume is summarized in the bottom sections of Tables 5 and 6. The mean FAO percent was lower in low volume trials (7%) than intermediate or high volume trials, where FAO percent was about 17%. The difference was significant, a 2 (training condition) x 3 (volume) mixed ANOVA indicated an effect of volume $F(2, 34) = 13.866$, $p < .001$, but no effect of training or their interaction. Report scores for FAO events followed the same trend, $F(2, 34) = 11.223$, $p < .001$.

Hit percent and FAO percent are summarized by density in Table 7, with report scores in Table 8. The top section summarizes results for incident events. Hit percent is highest on low density (0.91) trials, but otherwise seems to increase slightly with increasing density. A 2 (training) x 5 (density) mixed ANOVA revealed that the hit percent was affected by density ($F(4, 68) = 22.183$, $p < .001$), but not training, with no interaction between the two. Again, report scores for incident events show a similar pattern, although there seems to be greater difference between the two measures in the control condition than in the experimental condition. A 2 (training) x 5 (density) mixed ANOVA on report score for incident events revealed that report scores for incidents were affected by density ($F(4, 68) = 24.240$, $p < .001$), but not training, with no interaction between the two.

Table 7
Mean Proportion (SD) of Incident and FAO Events Submitted, by Density

Condition	Training Condition: Control	Training Condition: Experimental	Total
Incident Events			
Density			
3	0.91 (.15)	0.90 (.11)	0.91 (.12)
6	0.49 (.14)	0.57 (.16)	0.53 (.15)
9	0.50 (.19)	0.68 (.20)	0.60 (.21)
12	0.67 (.13)	0.65 (.21)	0.66 (.17)
14	0.76 (.22)	0.70 (.17)	0.73 (.19)
Total	0.70 (.13)	0.66 (.10)	
FAO Events			
Density			
3	0.08 (.07)	0.20 (.14)	0.14 (.13)
6	0.16 (.10)	0.17 (.11)	0.16 (.11)
9	0.03 (.06)	0.03 (.05)	0.03 (.05)
12	0.00 (.00)	0.00 (.00)	0.00 (.00)
14	0.17 (.22)	0.28 (.25)	0.22 (.23)
Total	0.09 (.08)	0.13 (.08)	

Table 8
Mean Report Score (SD) of Incident and FAO Events, by Density

Condition	Training Condition: Control	Training Condition: Experimental	Total
Incident Events			
Density			
3	0.83 (.16)	0.81 (.10)	0.82 (.13)
6	0.40 (.10)	0.48 (.13)	0.44 (.12)
9	0.43 (.18)	0.60 (.18)	0.52 (.20)
12	0.59 (.12)	0.63 (.21)	0.61 (.17)
14	0.61 (.20)	0.55 (.17)	0.58 (.18)
Total	0.57 (.11)	0.61 (.11)	
FAO Events			
Density			
3	0.07 (.06)	0.18 (.13)	0.13 (.11)
6	0.11 (.07)	0.12 (.08)	0.11 (.07)
9	0.02 (.04)	0.02 (.05)	0.02 (.04)
12	0.00 (.00)	0.00 (.00)	0.00 (.00)
14	0.14 (.17)	0.23 (.20)	0.18 (.19)
Total	0.07 (.06)	0.11 (.07)	

Data for FAO events across density is summarized in the bottom sections of Tables 7 and 8. The mean FAO percent was very low for the intermediate density levels of 9 and 12, but high for the highest density level of 14. A 2 (training condition) x 3 (volume) mixed ANOVA indicated an effect of density $F(4, 68) = 13.806$, $p < .001$, but no effect of training or their interaction. Reports scores for FAO events also varied by density, $F(4, 68) = 13.359$, $p < .001$, but not training condition.

Decision-making biases. Upon post-hoc assessment of some FAOs, the experimenters concluded that the composition of information items for some FAOs did not match the definition of the information traps, as described in training, accurately enough to allow valid scoring of the data associated with those FAOs. Thus, not all of the biases that participants were trained on were associated with FAO events within the experiment. None of the FAO events associated with the oversensitivity and sample size biases were deemed usable, and therefore are not discussed further in the results.

FAO percent was compared across bias type, with results presented in Table 9. There was no reliable difference in FAO percent between training conditions. FAOs associated with the vividness, absence of evidence, and availability biases were most likely to be reported, with little or no sensitivity to either randomness or persistence of discredited information. A 2 (training condition) x 5 (bias type) mixed ANOVA showed a significant effect of bias type $F(4, 68) = 7.99$, $p < .001$, but no effect of training condition or their interaction.

Table 9
False Alarm Opportunity Percent (SD) by Bias Type and Training

	Training Condition		
Bias Type	Control	Experimental	Total
Absence of Evidence	0.17 (0.35)	0.15 (0.24)	0.16 (0.29)
Availability	0.12 (0.15)	0.11 (0.10)	0.12 (0.12)
Persistence of Discredited Information	0.00 (0.00)	0.05 (0.10)	0.02 (0.07)
Randomness	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Vividness	0.19 (0.20)	0.30 (0.20)	0.25 (0.20)
Total	0.10 (0.11)	0.12 (0.10)	

Verbal reports of traps. Participants were asked to report to the experimenter any potential error opportunities (control group) or information traps (experimental group) they encountered, along with the associated items or "indicators" of the trap. Participants in the control group received credit for correctly identifying the associated bias if their verbal description of the error opportunity showed an understanding of the bias (i.e., they did not have to report the bias by the name provided to the experimental group).

Table 10 shows the proportion of FAOs that were identified as decision making traps (regardless of whether the correct bias was associated with the trap). Although performance was similar between training conditions with regard to FAO percent, participants in the control group were much less likely to provide verbal reports of traps. Within the experimental group, a trap was reported for 45% of persistence of discredited information events, compared with 20-30% for the other FAO events.

Table 10

Proportion of FAO Events Verbally Reported as a Trap (Any Bias)

	Training Condition	
Event Bias Type	Control	Experimental
Absence of Evidence	6%	30%
Availability	7%	24%
Persistence of Discredited Information	11%	45%
Randomness	11%	20%
Vividness	15%	26%
Total	10%	29%

Table 11 shows that the participants who were given bias training had mixed success in associating the correct bias to verbally reported information traps, reporting roughly a third of the absence of evidence and persistence of discredited information FAOs, but less than a tenth of the availability and vividness FAOs. Control group participants were generally less likely to provide verbal reports, and only rarely provided a correct description of the bias. Comparing these results to those presented in Table 10, experimental group participants provided verbal reports for availability, randomness, and vividness FAO events about 20-25% of the time, but only rarely associated the trap with the correct bias.

Table 11

Proportion of FAO Events Given Correct Verbal Report

	Training Condition	
Event Bias Type	Control	Experimental
Absence of Evidence	6% (6%)	30% (17%)
Availability	0%	5% (92%)
Persistence of Discredited Information	6% (50%)	33% (57%)
Randomness	4% (100%)	0%
Vividness	0%	9% (33%)
Total	3%	15%

Note. Numbers in parentheses indicate how often the trap was reported correctly.

Although it appears the experimental group was moderately successful in identifying the absence of evidence and persistence of discredited information traps, the results must be considered in the context of the mistakes that were made. The numbers in parentheses provide an indication of the precision with which participants applied their verbal reports by showing how often participants were correct with their verbal identification of a trap overall. Thus, although only 5% of availability traps were correctly identified with a verbal report, when an availability report was given, participants were correct 92% of the time. Although the absence of evidence trap was correctly identified 30% of the time, the overall success in reporting the bias was only 17%. In the latter case, verbal reports of an absence of evidence trap were common, but rarely correct.

Item modality. The proportion of data items included in reports of incidents is shown in Figure 7, with results separated by modality and training condition. The pattern of performance is quite similar in both conditions. Unlike the previous experiment, audio files were often submitted and were the most likely relevant items to be included in incident reports, followed by icon and text files. A 2 (training condition) x 6 (modality) mixed ANOVA for incident data item inclusion indicated an overall effect of modality $F(5, 85) = 4.47$, $p = .001$. Neither the main effect of training condition nor the interaction was significant.

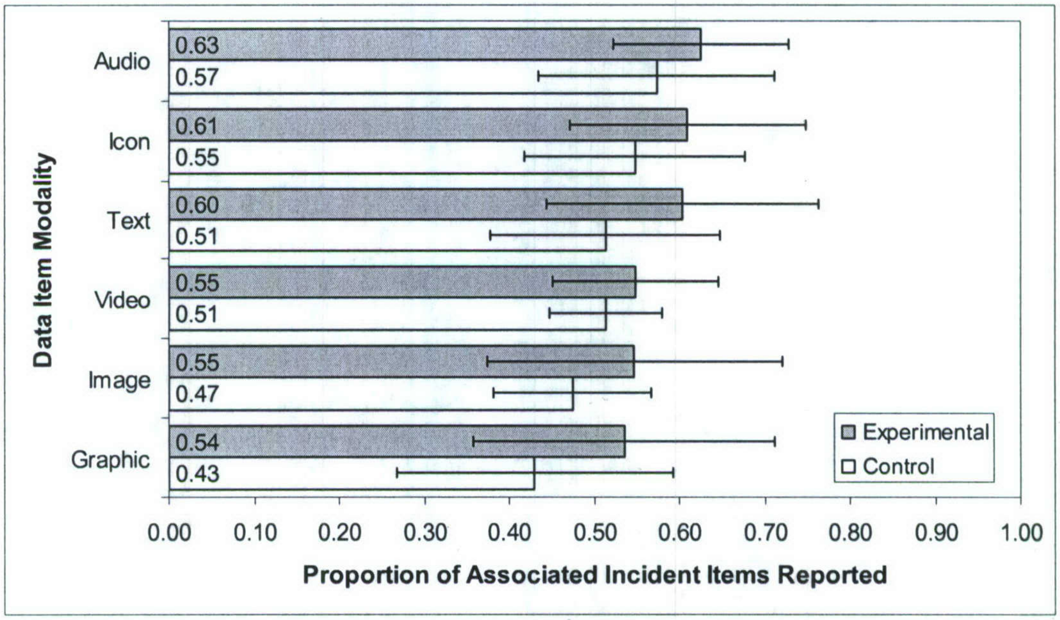

Figure 7. Proportion of Experiment 2 incident items submitted, by modality.

The proportion of data items included in reports associated with FAO events differed depending on the modality of the data item ($F(5, 85) = 14.10$, $p < .001$), but not training, as indicated by a 2 (training condition) x 3 (modality) mixed ANOVA. The proportion of submitted data items associated with FAO events is shown in Figure 8, across levels of training. Graphic and text files were less likely than other formats to be included in reports associated with FAOs.

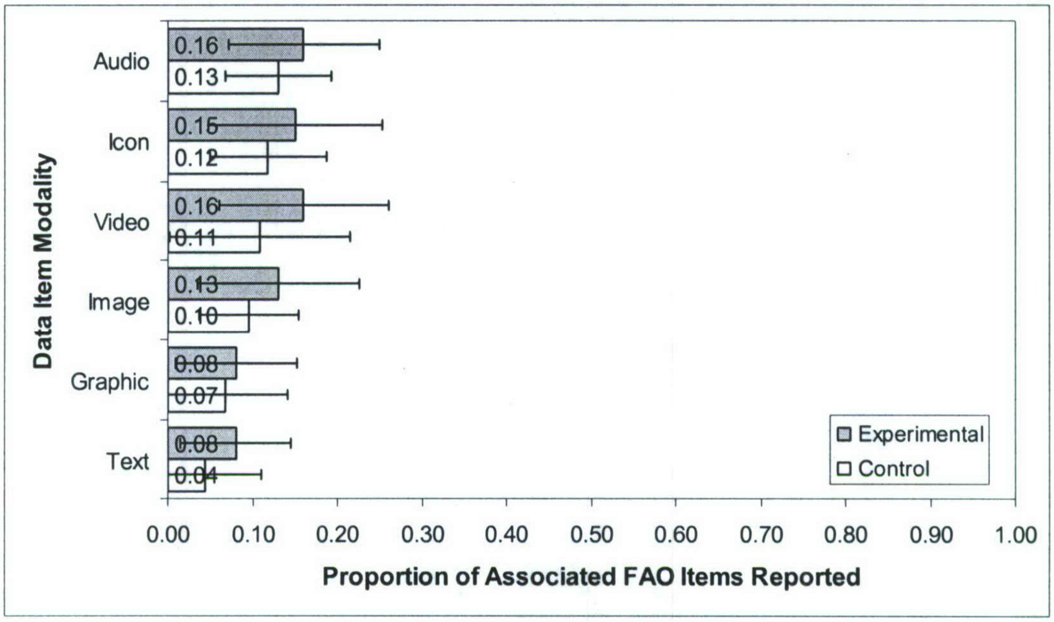

Figure 8. Proportion of Experiment 2 FAO items submitted, by modality.

Experiment 2 Discussion

Experiment overview. Experiment 2 extended the variables investigated in Experiment 1 to include the effect of training individuals about markers of the decision-making biases confirmed in Experiment 1 (experimental condition) vs. simply alerting them to the possibility of "error opportunities" (control condition).

Training. Of primary importance in Experiment 2 is the effect of training on FAO percent. Contrary to our hypothesis, participants who received anti-bias training performed no better than those in the control group, who were merely alerted to the possibility of error opportunities. Nonetheless, a comparison of reporting data with Experiment 1 seems to indicate an overall affect of alerting; that is, the tendency to commit decision making errors was much less in Experiment 2, with participants reporting just 14% of FAOs versus 35% observed in Experiment 1. Averaged over the five types of bias common across both experiments, the same trend was observed, with both experimental (12%) and control (10%) group participants reporting fewer FAOs than participants from Experiment 1 (34%). The hit rate was also less, with 63% in Experiment 2 versus 77% in Experiment 1. The latter finding is perhaps suggestive of a greater level of vigilance applied to the task. Participants may have been inclined to work more slowly to avoid error opportunities.

The only analysis that showed an effect of training was in the verbal reporting of bias conditions. Thus, the particular type of training received (experimental vs. control) did not seem to increase or reduce the tendency to submit incident reports of FAO events, and did not affect the ability to detect incidents. This was true when considering training in combination with volume and density, as well as decision biases and modality.

Verbal report of bias potential. As expected, participants who received anti-bias training provided more verbal reports of traps than those who did not receive the training. Nonetheless, although the experimental group verbally identified traps almost three times more often than the control group, there was no difference between groups in actual submissions of reports associated with FAO events. Two possible explanations include: (1) that experimental group participants sometimes recognized patterns of data items associated with traps, but decided to submit reports anyway; and (2) that both groups were equally likely to make decision errors, but the experimental group was better able (or more inclined) to provide verbal reports of traps for the events on which they did not make a reporting error. In either case, the training provided to the experimental group did not seem to provide an advantage over the alerting provided to the control group.

Given the rate of *correct* verbal reports of traps (i.e., traps that were identified and associated with the correct bias), it appears the training provided was only mildly effective in helping participants identify the markers for a given bias. It was only on events associated with the absence of evidence and persistence of discredited information biases that the experimental group was able to associate the correct bias with consistency (30% and 33% of the time, respectively). Further analysis indicated that much of the success in reporting the absence of evidence bias may not have been due to an understanding of how to recognize the markers for that bias. Quite the opposite, when reporting absence of evidence, the experimental group was correct just 17% of the time. This finding suggests that participants did not understand the bias, using it as a "catchall" category for reporting error opportunities. It seems likely that participants wrongly attributed this bias to a simple lack of a key indicator (a characteristic all FAOs had in common). This interpretation was inadvertently supported by misleading language used in the summary sheet describing the bias, which in part stated that an absence of evidence trap occurs when "a critical piece of information you should have to report an incident with confidence is missing." In fact, the bias occurs when people make decisions based on the presence of information (broken window, kids playing baseball outside) without considering the absence of information that disconfirms an explanation (no glass inside the house, no ball inside the house).

To summarize, the verbal report data show that the experimental group was able to report traps with greater frequency, but had a very difficult time associating the correct bias to the trap. This result seems in part due to a misinterpretation of the absence of evidence bias, which was associated with the majority of traps that were reported. The experimental group did show some success with the persistence of discredited information bias, associating the bias correctly 57% of the time and identifying 33% of the relevant events. Nonetheless, there was no evidence overall that the anti-bias training reduced the likelihood of committing characteristic errors over simple alerting to error opportunities. It appears likely that the anti-bias training provided to the experimental group was inadequate, as the group was unable to correctly associate biases with the traps that were verbally identified. As such, the training provided may have simply amounted to alerting, in that the experimental group did not seem to absorb the training at a level that enabled the transfer of that knowledge to the task.

Training, volume, and density. There was no effect of volume on reporting incidents. As in previous studies, this may suggest that density and volume levels implemented in this experiment were not high enough to affect incident performance. Volume did affect the reporting of FAOs, however. Participants were particularly unlikely to fall for FAOs when the volume was lowest, and more likely to fall for FAOs at higher volumes. Unlike Experiment 1, increasing the amount of irrelevant information increased the likelihood of participants deciding that an ambiguous pattern of information is indeed reportable. In this case, participants were more likely to just say something happened as the extraneous information increased. They mistakenly submit reports more often, rather than just holding the information because they are not sure. The reason for the difference between experiments with respect to volume is difficult to ascertain.

Density affected reporting of incidents. Specifically, having the lowest amount of relevant information yielded the best performance for detecting incidents, and increasing beyond this amount dramatically reduced performance. This is the first instance of information quantity affecting the reporting of incidents. Because information quantity levels were similar to previous experiments, this suggests that these incidents were slightly easier to detect than others. The effect is again in the expected direction, causing more reports to be mistakenly submitted.

Participants were most susceptible to FAOs when there was the least or the most amount of relevant information available. One possible explanation for this peculiar finding is that they adopt one strategy when increasing relevant information to a certain point, and then adopt another when increasing it even more. The first strategy is the strategy implied in Experiment 1, where the lower FAO percent and report score is due to them not having the resources to make the final decisions about something they are aware of but not sure about. The second strategy is the strategy implied in other aspects of this research, where the higher FAO percent and report score is due to being so overwhelmed with information that the participants just say it is there if they're not sure. Speculating why this shift occurs, perhaps participants do not make the final decision and submit the FAO at medium densities because they think they can get back to the report to make a final decision. But when the density increases enough, they are so overwhelmed that they adopt a "now or never" view of dealing with the information, needing to make a decision on the spot. Thus, it seems there is an optimal amount of relevant information for the least susceptibility to FAOs. If users are busy enough, they will not have enough resources to go back and evaluate the information, and thus will not submit the report and fall for the FAO. If users are too busy, they will just submit the report right away if they're not sure about the status of the event.

Decision-making biases. FAO events with the vividness bias were the most likely to be submitted; absence of evidence and availability were less likely, and persistence and randomness hardly ever occurred. Although fewer biases were studied in Experiment 2, the overall trend was similar to what was observed in Experiment 1. The prevalence of the vividness bias again suggests that direct subjective communication of information makes people think that an event has occurred, regardless of the lack of essential information ensuring an event indeed occurred. The slight susceptibility to the absence of evidence bias suggests that given an unclear suggestion of a military event, the fact that evidence is missing may not be given its due consideration. The slight susceptibility to the availability bias suggests that decisions may be made based on what the most recent, convenient, or otherwise accessible explanation is for a somewhat ambiguous pattern of information.

Item modality. Despite slight differences and many comparisons, there are again some clear trends in the modality of items attached to reports. Regardless of event type, audio and icons were included consistently more often, and graphs included consistently less often. For all items associated with incidents, the inclusion rate for graphics was especially low relative to other types of modalities. As suggested in the Experiment 1 discussion section, the ambiguous nature of interpreting graphs may have caused them to not be attached to incidents, because they did not provide concrete enough evidence to be considered as indication of an event, relative to other available sources of information. When considering only the essential items associated with incidents, the pattern of results did not change; no type of modality was favored for inclusion as the key or confirming indicator.

The proportion of items included in FAOs was, overall, much lower than in incidents, which is consistent with the lower report score associated with FAOs. Graphic and text files were less likely than other formats to be included in reports associated with FAO opportunities. Although the low proportion of graphs is consistent with other event types (but not other experiments), text messages were less important for these FAO events than they were for incidents. While graphic files can be ambiguous and are somewhat open to interpretation, text messages are less open to interpretation and more likely to clearly indicate the information intended to be conveyed. The ambiguity of graphic files and the literal nature of text files make it unclear whether the degree of ambiguity has anything to do with why certain data items are included more than others, as suggested in earlier experiments.

Experiment 3

The purpose of Experiment 3 was to determine whether the seven characteristic errors observed in Experiment 1 operate in the context of team decision making. The experiment was designed to assess the effects of data format on assignment of meaning with the types of errors and biases observed in these tasks when performed by small teams. The pattern of susceptibilities to biases may change extensively when considering communication and social dynamics in a small group of people. Based on the results of Experiments 1 and 2, density and volume were not considered in Experiment 3.

Experiment 3 Method

Participants. Participants for Experiment 3 were 15 volunteers recruited from the student body at Georgia Tech. Five of the 15 participants were ROTC students, and the remainder were regular undergraduate students. None had previous experience on the task.

Participants were divided into five teams, made up of three people each. One team member was designated as leader and had the final decision about report submissions. The team leader was designated to be someone who stood out as a good leader to the experimenter. A "good leader" was experimentally defined as someone who was interactive and assertive or clearly understood the task better than others in the group. If no one was an obvious candidate according to these criteria, the experimenter asked for a volunteer from the group. If no one volunteered, the experimenter randomly chose a leader. The remaining two team members were designated "Support 1" and "Support 2," and were primarily responsible for processing their assigned data items.

Apparatus. Three-way voice communication was added to the general apparatus previously described. Participants used headphones with attached microphones that, when sent through a computerized communication system, enabled participants to converse and utilize the AID without disturbing the other participants.

For team trials, the three workstations were linked. The AID presented data items into three separate folders corresponding to the appropriate team member (Leader, Support 1 or Support 2). All three folders were visible and accessible on each instance of the AID, although the default folder was set to the appropriate workstation (e.g., the participant filling the Support 1 role could view incoming messages in the Leader and Support 1 folders, but the default folder was Support 1).

Procedure. The experiment was divided into four sessions. The first two sessions were performed individually, and the last two sessions were performed in teams of three. In the first session, participants were oriented to the task and interface, with explanations of the incident categories and types and a thorough introduction to the AID. In the second session, participants performed practice trials individually to aid in familiarization with the task and incident types. No anti-bias training was administered in this experiment. The individual training sessions took a total of 5 hours.

In the third session, participants received group training on how to function as a unit before performing additional practice trials. This was typically given the same day as, and immediately preceding, the test session and allowed team members to familiarize themselves with each other and to begin the process of functioning as a group. In the final session, team performance was assessed on two test trials. The final session ran for 2 hours.

The basic task was similar to that in Experiments 1 and 2, but occurred in the context of a team. Members worked together to process a large number of data items in rapid succession and to make a decision about whether to report the presence or occurrence of target incidents. Team members determined which pieces of information were relevant to the mission and were asked to file reports based on this information.

The task varied somewhat for the participant designated as the Leader. Whereas the two support personnel received a scheduled delivery of data items to process (just as in the individual trials), the team leader received only "prompts." Prompts were brief text messages that relayed intelligence to the team leader on an event that might occur (e.g., "We have reason to believe that there may be a sniper attack in Midland today."). The Leader's task was to confirm or deny the event based on the information processed by the team. Participants were instructed to confirm the prompt if there was relevant information sufficient to file an incident report. If there was insufficient information to file a report, participants were instructed to deny the prompt. Confirmation of a prompt required a report submission.

There were relatively few prompts in a given trial. Most of the time, the Leader was either verifying information in a report submitted by the support personnel or helping to process items by accessing one of the two folders assigned to the support personnel. The leader was primarily responsible for final submission of all reports, in addition to responding to all prompts. Support personnel received the bulk of the data items and tended to make the initial judgment of relevance.

Successful completion of the task required teamwork. Data items associated with incident and FAO events were often split between inboxes; thus, it was often necessary for team members to communicate to determine whether or not they had received information relevant to a particular event. Participants also could examine the contents of folders created by their team members and add additional information as required.

Stimuli. A new set of stimuli was created for Experiment 3. The stimulus set was based on the same 16 incident events used in previous experiments, with the same criteria for an incident occurrence. Although many of the data items were reused from previous experiments, almost all were modified and sorted into incident and FAO events unique to this experiment.

In addition to the incident and FAO event types used in the two previous experiments, Experiment 3 included a number of false alarm control (FAC) events. FAC events were constructed to be as similar to FAO events as possible, but were altered to eliminate sensitivity to the associated bias. For example, the oversensitivity bias occurs when several pieces of identical information are interpreted as providing additional or corroborating evidence. An oversensitivity FAO for a sniper attack might consist of several reports of a suspicious man on a roof, but all from the same source (e.g. a report from an observer, a news report based on an interview with the observer, and a police report derived from discussions with the observer). The corresponding FAC for a sniper attack would consist of the same information, but coming from different sources.

In creating controls for the characteristic errors under study, we hoped to be able to show FAO percents significantly higher than the corresponding FAC percents as further proof that a false alarm is due to the underlying bias for which it is designed. Unfortunately, it was determined after the reserach that the corresponding FAC for the randomness and oversensitivity biases could be reasonably construed as true incidents, negating the purpose of the control. Further, the removal of the error marker for the absence of evidence bias effectively turned it into an incident as well. Thus, only the FACs associated with the availability, vividness, persistence of discredited information, and sample size remained false alarms.

The test session in Experiment 3 assessed performance on two longer trials rather than multiple shorter trials of varying density and volume. Participants in the support roles received a small number of items at the beginning of the trial, followed by two new items every 30 seconds. Across both trials, there was one incident event corresponding to each of the 16 incident types identified in training (eight in each trial). Trial 1 contained one FAO for each of the seven biases under study. Trial 2 contained one FAO and one FAC for each of the seven biases under study. The FAC events in Trial 2 corresponded to the FAO events in Trial 1, making seven FAO/FAC pairs. There were eight prompts in Trial 1 and twelve in Trial 2—all went to the Leader.

For both test trials, two incident events went to Support 1 only, two went to Support 2 only, and the remaining four were split between Support 1 and Support 2. FAOs and FACs were similarly split. For both trials, two FAO events went to Support 1 only and two went to Support 2 only, with three split between Support 1 and Support 2. FAC events were similarly split for Trial 2 only. Filler items were split between Support 1 and Support 2 such that each received an equal number of items, and an equal number of each item modality.

Characteristics of the filler item were another aspect of the information being delivered to the team that was considered. The filler items prepared for Experiment 3 were classified into six categories:

- *Superficial Similarity:* These data items share superficial attributes with an event, but their content is irrelevant to the incident. For example, a text message might share a key word with an incident, even though the content of the message is unrelated. We hypothesize that these stimuli are erroneously considered as related and thus are more likely than standard filler to be included with a report.
- *Sensationalistic:* These data items contain heavily exaggerated claims or threats. We hypothesize that the sensationalistic aspect of these stimuli causes these data items to be more likely than standard filler to be included with a report.
- *Irrelevant:* These data items are completely irrelevant to any sort of terrorist event. We hypothesize that these data items should be nearly immediately dismissible as not relevant, and thus less likely to be included with reports.
- *Oddball:* These are stimuli that are extremely rare or unique. We hypothesize that more attention will be paid to these data items, and thus they will be more likely to be included with a report than conventional filler.
- *Over-Detailed:* These data items include a large amount of extraneous information. We hypothesize that participants will erroneously consider this information to be important because of the level of detail, and thus be more likely to include it with a report than conventional filler.

Experiment 3 Results

Event type. The mean proportion of events reported was highest for incident events (M = 0.68, SD = 0.03), followed by FAC events (M = 0.46, SD = 0.23) and FAO events (M = 0.21, SD = 0.14). The difference between hit percent and FAO percent was statistically significant, ($F(1, 4) = 57.51, p = .002$). Mean report scores followed the same trend for incidents (M = 0.50, SD = 0.07), FACs (M = 0.31, SD = 0.18) and FAOs (M = 0.16, SD = 0.09), and the difference between report scores for incident events and FAO events was again significant, ($F(1, 4) = 30.96, p = .005$).

Decision biases. The FAO percent for each bias type is shown in Table 12, where higher scores represent a greater tendency to submit a report of an incident that did not occur and thus a greater susceptibility to the bias. FAO percent was very high for the oversensitivity bias (0.70), especially given that the overall proportion of FAO events submitted was just 0.21. Individuals in Experiment 3 also showed the highest sensitivity to the oversensitivity bias. The low sensitivity to FAOs associated with the persistence of discredited info and sample size biases in this experiment is consistent with results from the previous two experiments. A one-way ANOVA for bias type in Experiment 3 revealed a significant effect on FAO percent, $F(6, 24) = 5.610, p = .001$).

Table 12
False Alarm Opportunity Percent by Bias Type

Bias Type	M	SD
Availability	0.20	0.27
Absence of Evidence	0.10	0.22
Oversensitivity	0.70	0.27
Persistence of Discredited Information	0.10	0.22
Randomness	0.20	0.27
Sample Size	0.00	0.00
Vividness	0.20	0.27

Bias and event type. For each of the seven biases, an FAO was paired with its corresponding FAC. The proportion of events reported of each type is shown in Table 13. Although the original intent of including FAC events was to provide additional false alarm opportunities that did not correspond to any of the decision biases under study, the resulting FACs associated with the randomness, absence of evidence, and oversensitivity controls were reportable incidents rather than false alarms. On the remaining four pairs (availability, vividness, persistence of discredited information, and sample size), the FAC events were simply false alarm events without the associated decision bias. FAC percent was similar to FAO percent on the persistence of discredited information and sample size pairs, but actually higher for the availability and vividness pairs where the bias was removed. Interestingly, teams were more likely to report the vividness control with the vivid material removed. Nonetheless, alpha-adjusted paired-samples t-tests showed that none of the differences were significant. Overall, few teams reported FAO events, with the exception of the event with the oversensitivity bias.

Table 13
FAO and FAC Percent by Bias Type (Paired Events Only)

	Event Type	
Bias Type	FAO	FAC
Absence of Evidence	0.20 (0.45)	0.60 (0.55)
Availability	0.20 (0.45)	0.60 (0.55)
Oversensitivity	0.60 (0.55)	0.60 (0.55)
Persistence of Discredited Information	0.00 (0.00)	0.20 (0.45)
Randomness	0.20 (0.45)	0.60 (0.55)
Sample Size	0.00 (0.00)	0.00 (0.00)
Vividness	0.40 (0.55)	0.60 (0.55)

Item modality. The proportion of reported data items associated with incident events is shown in Figure 9, with results separated by item modality. Overall, associated video and audio items were most likely to be included in incident reports. Text and image items were the least likely to be included. A one-way repeated measures ANOVA showed a significant effect of modality in reporting items for incident events, $F(5, 20) = 4.86, p = .005$.

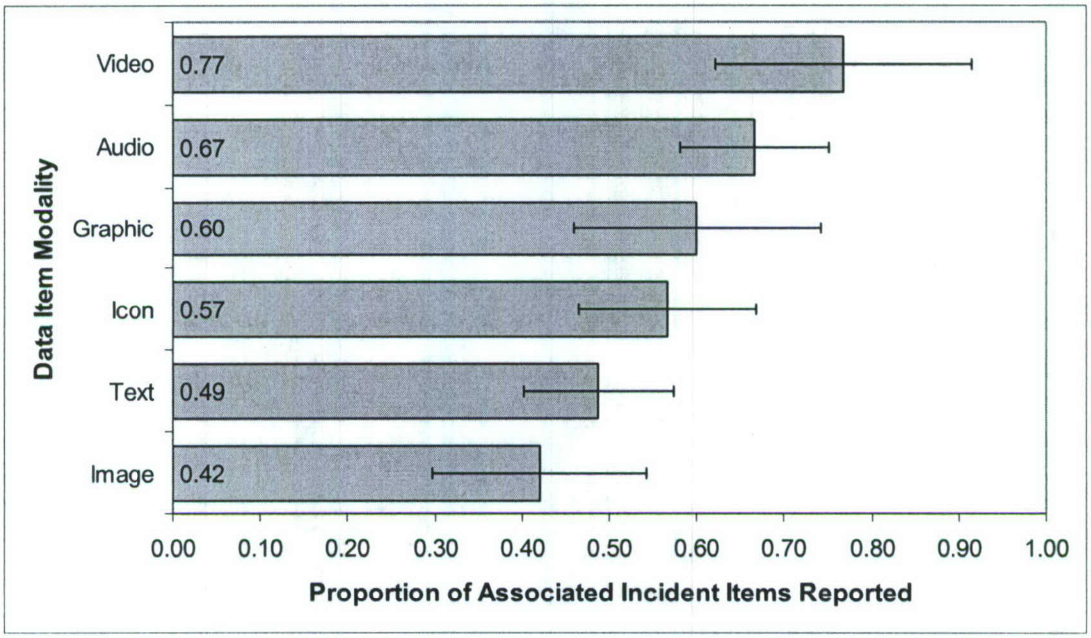

Figure 9. Proportion of Experiment 3 incident items submitted, by modality.

The proportion of reported data items associated with FAO events is shown in Figure 10, with results separated by modality. The rank order of reporting items associated with FAO events is almost the reverse of what was observed with incident events. Whereas important image and icon items were among the least likely to be included in reports of incident events, they were most likely to be used in reporting FAOs. A one-way repeated measures ANOVA also showed a significant effect of modality in reporting items for FAO events, $F(5, 20) = 5.99, p = .002$. Except for a lower tendency to report graphic items, the rank order of item format is consistent with Experiment 1, although the reliability of the observation is questionable due to the high variability of the data.

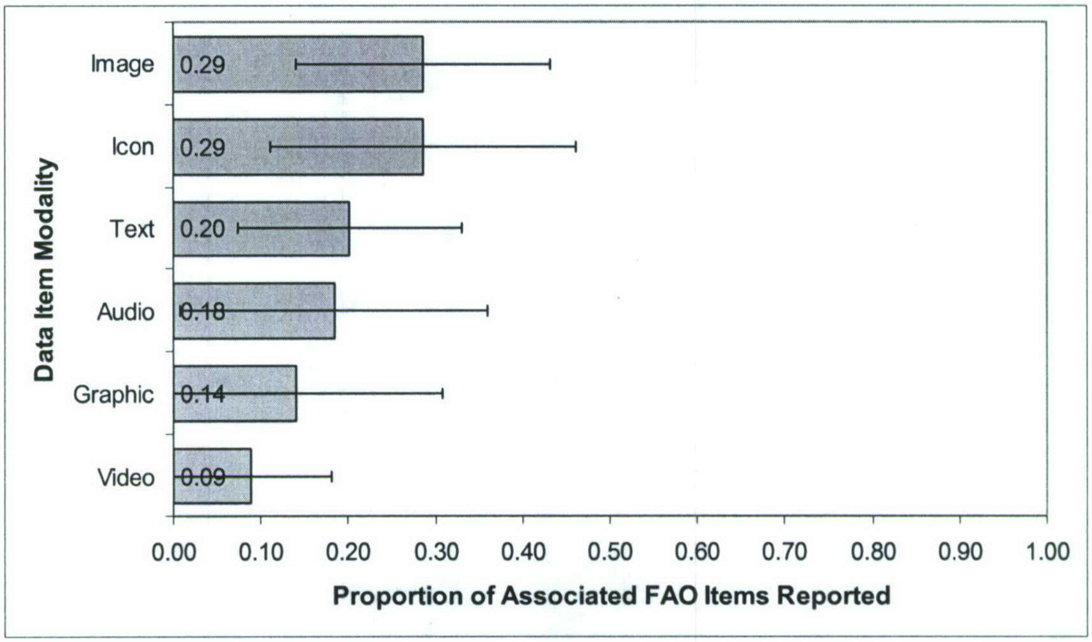

Figure 10. Proportion of Experiment 3 FAO items submitted, by modality.

Response to prompts. On average, teams responded correctly to about half of the prompts given (M = 0.56, SD = 0.08). Of the prompts that were answered, 64% were answered correctly, on average (SD = 0.11). Performance varied greatly depending on the correct answer for the prompt. Prompts that were to be confirmed were answered correctly 80% of the time, but prompts that were to be denied were answered correctly only 32% of the time, $F(1, 4) = 17.19$, $p = .014$.

Table 14 compares the mean submission rate of events that were associated with prompts versus those that were not. For each type of event, submission rates were higher when the event was associated with a prompt. Only the FAO difference was statistically significant, $F(1, 4) = 9.89$, $p = .035$. Teams were more likely to report FAO events as incidents when the events were associated with a prompt.

Table 14
Report Submission by Association With Prompt

	Mean Proportion Submitted (SD)	
Event Type	Events Associated With Prompts	Events Not Associated With Prompts
Incident	.70 (.07)	.63 (.14)
FAO	.37 (.22)	.10 (.10)
FAC	.60 (.38)	.27 (.28)

Extra attachments. Over one third of reports filed by teams contained at least one incorrect item attachment (M = 0.36, SD = 0.12). Extra attachments were more common in submitted FAO reports (M = 0.40, SD = 0.20) than incident reports (M = 0.33, SD = 0.21), though the difference was not statistically significant. The number of extra attachments in a report ranged from 1 to 8. The extra items were evenly split between items associated with other events (51%) and filler items (49%).

Incorrect items that were included with a report but were intended for another event tended to be either items for a different event of the same incident category (e.g., two separate sniper events reported as a single incident, 48%), or events of a very similar nature (e.g., items for a road mine event filed with a car bomb report, 42%).

Of all filler items attached to reports, 34% were classified as showing superficial similarity to the event, 22% were classified as sensationalistic, 22% were classified simply as standard filler, and 13% were classified as oddball filler (see Table 15). Taken together, people were just as likely to include in their reports the stranger sensationalistic and oddball filler items (35%) as the superficial similarity items.

Table 15
Classification of Filler Items Filed with Submitted Reports

Filler Type	Proportion of All Filler Items Submitted
Superficial Similarity	34%
Sensationalistic	22%
Standard	22%
Oddball	13%
Over-Detailed	6%
Irrelevant	3%

Experiment 3 Discussion

Experiment overview. The primary goal of Experiment 3 was to examine the role that cognitive biases play when teams are performing the rapid decision making task. A new type of false alarm was implemented in Experiment 3, a false alarm control (FAC), which offered the same collection of information as a FAO, but without the bias implemented. Experiment 3 also considered aspects of irrelevant data items (e.g., filler) that may cause relevance to be attributed.

Event type. For participants working in teams, the hit percent (0.68) and FAO percent (0.21) were more comparable to participants in Experiment 2 (0.63 and 0.14, respectively) than participants in Experiment 1 (0.77 and 0.35, respectively). The reason for the difference in performance between Experiments 1 and 3 again may be the result of increased vigilance. Whereas in Experiment 2 an increased level of vigilance might be predicted from being alerted to error opportunities, in Experiment 3 an increased level of vigilance might occur as a result of performing the task as a team. Fewer false alarms are expected given a second participant to check each report before it is submitted.

Controls for the FAOs were added in Experiment 3 to help make the case that the errors made in these studies are characteristic of the biases identified in previous research. Based on our assumption that errors would be driven by characteristic biases, we would predict the FAO percent to be higher than the FAC percent. This did not occur; in fact, the opposite trend was observed, with participants reporting FAC events more often than FAO events. Although this trend is worth noting, the difference was not statistically significant, especially given that the comparisons provided represent differences between a single FAC/FAO pair for each bias. To provide the statistical power necessary to reach reliable conclusions, further research might examine this more extensively, with multiple FAC/FAO pairs for each bias.

As mentioned in the explanation of the stimuli for Experiment 3, it was determined after data collection that the FAC events for the oversensitivity and randomness biases could reasonably be construed as true incidents. Although those events were modified slightly (fixed) for Experiment 4, it was only after data collection was complete for both experiments that a problem with the FAC for the availability bias was discovered. An error associated with the availability bias occurs when information is falsely interpreted in light of a convenient explanation (e.g., if a rash of suicide bombings has occurred recently, an explosion in a market might be deemed a terror attack although no real evidence exists to support the assertion). The FAC for the availability bias has the same information, but no item to create the "convenient explanation" (i.e., no information on a rash of suicide bombings). Although the stimulus materials were appropriate for the control in Experiments 3 and 4, the event was randomly assigned to be paired with a prompt item that effectively replaced the convenient explanation (e.g., "This is headquarters, we have received information on a planned suicide bombing today, please confirm or deny."). Thus, when paired with a prompt, the availability FAC essentially reverted to an FAO.

Decision biases. The overall trend in FAO percent was similar to the trends observed in previous experiments. Both individuals in Experiment 1 and teams in Experiment 3 were observed to show the greatest susceptibility to the oversensitivity bias and low susceptibility to the sample size bias (not examined in Experiment 2). There seems to be only scant evidence of the persistence of discredited information bias in this rapid decision making task, with participants in all three experiments showing little or even no sensitivity.

Item modality. Again, trends in item modality were observed. Incident reports contained higher proportions of audio and video items, which may be more memorable and therefore more likely to be included even when team members are asking for information. Text and image items were much less likely to be included in incident reports than other modalities. Although text is a less vivid description of information, images allow the participant to see the information with their own eyes. Once again, these conflicting characteristics of the modality make it unclear what about these modalities cause the different inclusion rates.

The rank order of reporting items associated with FAO events is approximately the reverse of what was observed with incidents. Unlike the trend in incidents, image and icon items were the most likely to be used in reporting FAOs. Although the interpretation of the particular pattern is not clear, the fact that the reverse trend occurred indicates that there may be some consistent aspects of the different modalities causing some to be valued higher than others, depending on the type of event. Finally, there were no significant differences between modalities in the proportion of reported data items associated with FAC events.

It is difficult to explain why trends in the current experiment differ from those observed in Experiments 1 and 2. Where incident text items were reported 84% of the time in Experiment 1, they were reported just 49% of the time in Experiment 3. Incident image items also were reported in higher proportions in Experiment 1. There are also quite different trends in item modality associated with FAO events, although the very high variability in the data for Experiments 2 and 3 make such comparisons questionable. These noteworthy differences in item modality trends across experiments suggest that the differences in reporting may be driven by factors other than data format.

Response to prompts. Participants had a much higher report score when FAOs were associated with prompts than when they were not. This may be due to the indecisiveness associated with FAOs: The suggestive nature of prompts may have made participants submit a report they were not sure about or remind them of a report they were indecisive about.

The discrepancy suggests the prompts influenced teams to view FAO events more favorably than they might otherwise, in effect introducing an availability bias through a different mechanism.

Percent of reports containing extra attachments. More than one third of reports filed by teams contained at least one incorrect item attachment. Thus, although extra attachments were not uncommon, most of the time there were no extra attachments. In addition, there was no difference in the mean extra attachments between reports submitted with FAOs and with incidents. Together, these findings suggest that differences in responding to incidents and FAOs are due more to a difference in how relevant information is treated than to how irrelevant information is treated. Those items that were irrelevant to the event, however, were not completely unrelated: Most extra attachments were somehow superficially related to the event reported, and participants were particularly likely to assume that unrelated pieces of information (filler) with superficially similar cues were relevant. In this way, the extra attachments that occurred were not random errors; they were deliberate selections that appeared to be relevant.

Experiment 4

Experiment 3 confirmed that the characteristic errors in decision making observed in Experiment 1 also are observed when decision makers work in small teams. Experiment 4 evaluated the effect of anti-bias training in teams by comparing the test trial performance of Experiment 3 participants with a new set of participants, all of whom were given a revised version of the training used in Experiment 2. The experiment was designed to assess the effects of anti-bias training and data format on assignment of meaning and on the types of errors and biases observed in these tasks when performed by small teams.

Experiment 4 Method

Participants. Participants for Experiment 4 were 15 volunteers, the majority of whom were ROTC students at Georgia Tech. Four participants were working professionals between the ages of 29 and 35. None had previous experience on the task. The participants were divided into five teams of three each. The same distribution of responsibilities in Experiment 3 applied to this experiment, with a Leader and two team members in "support" roles.

Training. All participants in Experiment 4 received anti-bias training similar to that provided for the experimental group participants in Experiment 2. An objective of the current experiment was to improve the training provided in Experiment 2. To that end, the number of biases for which participants received training was reduced from seven to five. This reduction applied only to bias *training*: FAO stimulus items for Experiment 4 remained identical to those provided in Experiment 3 (i.e., with FAO events associated with all seven biases).

Based on an examination of the results for previous experiments, especially Experiment 2, several changes were made. Due to an inherent similarity between the availability and randomness biases, the two were addressed in training as a single trap, "jump to conclusion." Both of these biases occur when people make judgments in light of a perceived relationship that is not supported by the data. In consolidating the training, it was assumed that participants would be able to identify traps associated with either bias, without needing to identify whether the perceived relationship was due to an available explanation brought on by recent events (availability), or one created in the mind to explain random data (randomness).

The absence of evidence bias was removed altogether from training. This bias was considered by the experimenters to be the most difficult to understand. Previous training also seemed to be misleading. In Experiment 2, it was used extensively in verbal reports of traps, but was most often used in an improper context. Although training could have been improved, part of the choice for eliminating this bias was to reduce the overall complexity of training. Because absence of evidence FAOs existed in the dataset, a general bias category was added to the training to provide a means for reporting those traps (or any other perceived traps that teams could not associate with their training).

As in Experiment 2, participants in Experiment 4 had access to a summary of the information traps discussed in training. The following list describes each information trap as summarized for participants in their training manuals. The bias(es) associated with each trap (in parentheses) were not provided to participants, and are included here to assist the reader:

- Jump to Conclusion (Availability/Randomness):
    - Definition: This occurs when information that seems to fit a convenient explanation is assumed to support that explanation, though no direct evidence exists to support that conclusion. People tend to interpret information based on what they can easily imagine or recall, and may have difficulty recognizing alternative explanations.

- Indicator: The indicator for this trap is the presence of information that matches recent events, but lacks the necessary evidence to prove that the individual items are interrelated, or related to that event.
- Redundancy (Oversensitivity):
  - Definition: This occurs when multiple pieces of information are believed to provide considerable evidence for an event, when in fact all the separate pieces of information were derived from the same source.
  - Indicator: The indicator for this trap is the presence of several messages that originate from the same source.
- Stubbornness (Persistence of Discredited Information):
  - Definition: This occurs when a belief persists despite the discovery of new information that discredits that belief. People may ignore contrary evidence, especially when their belief provides a convenient explanation for a set of circumstances.
  - Indicator: The indicator for this trap is a body of evidence that suggests a particular interpretation for an event, followed by a subsequent piece of information that disproves that interpretation.
- Small Sample (Sample Size):
  - Definition: This occurs when a report is filed using evidence from small sample sizes that contradicts evidence from larger sample sizes. Sometimes small samples are treated as equally important as large samples, though the evidence they provide is not as strong as that from a large sample. Weighing both groups the same (attributing equal importance to both) can result in an error in judgment, particularly if the small sample is inadequate for drawing conclusions.
  - Indicator: The indicator for this trap is the presence of conflicting information from groups of different sizes.
- Seeing (or Hearing) is Believing (Vividness):
  - Definition: This occurs because information that is vivid, concrete, and personal usually has a greater impact on decision making than abstract information. For example, because you can visualize an event better through video, audio, and/or image messages, these messages may influence the perception of an event more than a graph or text, even though the information may be inconclusive in and of itself. Similarly, if a report is received directly from an individual as opposed to one reporting on someone else's observations, the direct report may be a stronger influence in your reporting.
  - Indicator: The indicator for this trap is the presence of many related video, audio, and/or image messages, especially messages that are striking or dramatic.
- Other Error Opportunities (General)
  - Definition: Any other ambiguity or situation that may present an opportunity for error that does not match one of the other traps on which you have been trained.
  - Indicator: The indicator for this trap is any other ambiguous or misleading information not covered in the other trap categories that you believe might be mistaken as evidence of a reportable incident.

Further refinements to the training were made. Where necessary, descriptions of the biases were tailored to be more relevant to the context of the task, rather than focusing on a more abstract explanation of the bias. Examples were modified for clarity and some were added, where necessary. The practice trial following each bias explanation and example was simplified to include only the collection of information demonstrating the bias, and this information was revised to be a more lucid representation of the bias. Follow-up explanations of the bias indicators in the practice trial were reviewed at the end of each practice trial.

Procedure. As in Experiment 3, all participants received individual training, in two sessions spread across two days, in addition to a group training session that occurred immediately prior to the test session. The anti-bias training described above was provided in the second training session, in addition to (or in place of) the practice trials performed by participants in Experiment 3.

The team task was the same as in Experiment 3. Information elements handling was the same as in all previous experiments. The information flow to the teams was controlled in the same way as Experiment 3, on a predetermined sporadic schedule. Data item modality was balanced across team members. Reports were submitted in the same fashion as Experiment 1. As in Experiment 3, prompts were delivered during trials to the team Leader, who was instructed to confirm the prompt with evidence or deny the prompt.

Conditions and stimuli. Other than the training trials added to support anti-bias training in Experiment 4, the stimulus materials used were identical to those used in Experiment 3.

Experiment 4 Results

Event type. The mean proportion of events reported was highest for incident events (M = 0.48, SD = 0.123), followed by FAC events (M = 0.11, SD = 0.06) and FAO events (M = 0.03, SD = 0.06). Almost no FAO or FAC reports were filed by teams receiving bias training. The difference between hit percent and FAO percent was statistically significant, ($F(1, 4) = 152.743$, $p < .001$). Mean report scores followed the same trend for incidents (M = 0.39, SD = 0.12), FACs (M = 0.03, SD = 0.06) and FAOs (M = 0.05, SD = 0.03), and the difference between report scores for incident events and FAO events was again significant ($F(1, 4) = 87.57$, $p = .001$).

Overall, teams in Experiment 4 that received bias training reported fewer events of all types than teams in Experiment 3. Of greatest importance is the significant difference in the proportion of FAOs reported: ($F(1, 8) = 7.65$, $p = .024$). Because teams in this experiment received anti-bias training, they were predicted to submit fewer reports of FAOs than participants in the previous experiment. The difference in proportion of events reported was statistically significant for incidents ($F(1, 8) = 14.42$, $p = .005$) and FACs ($F(1, 8) = 10.13$, $p = .013$).

Decision biases. The FAO percent for each bias type is provided in Table 16. Although only a very small proportion of FAO events were reported, the events for which these occurred were associated with biases that produced higher proportions of false alarms in other experiments. The oversensitivity bias was associated with the highest FAO percents in Experiments 1 and 3. A 2 (training) x 7 (bias type) repeated measures ANOVA on the proportion of FAO events reported indicated a significant interaction, $F(6, 48) = 3.423$, $p = .007$. The main effects for training, $F(1, 8) = 6.70, p = .032$, and bias type, $F(6, 48) = 10.13, p < .001$, were also significant. Teams that received no training were more likely to submit reports of FAO events. This was largely due to an increased propensity of Experiment 3 teams to submit reports of FAO events associated with the oversensitivity bias.

Table 16
False Alarm Opportunity Percent by Bias Type

Bias Type	M	SD
Availability	0.20	0.27
Oversensitivity	0.10	0.22
Absence of Evidence	0.00	0.00
Persistence of Discredited Information	0.00	0.00
Randomness	0.00	0.00
Sample Size	0.00	0.00
Vividness	0.00	0.00

Bias and event type. For each of the seven biases, an FAO event was paired with a corresponding FAC event. The proportion of events reported of each type is shown in Table 17. Of the paired events, only the availability and oversensitivity events were reported, with one team reporting the availability FAO versus three teams reporting the availability control.

Alpha-adjusted paired-samples t-tests for the availability and oversensitivity biases showed no significant differences between FAO and FAC percent of corresponding events.

Table 17
FAO and FAC Percent by Bias Type

	Event Type	
Bias Type	FAO	FAC
Absence of Evidence	0.00 (0.00)	0.00 (0.00)
Availability	0.20 (0.45)	0.60 (0.55)
Oversensitivity	0.00 (0.00)	0.20 (0.45)
Persistence of Discredited Information	0.00 (0.00)	0.00 (0.00)
Randomness	0.00 (0.00)	0.00 (0.00)
Sample Size	0.00 (0.00)	0.00 (0.00)
Vividness	0.00 (0.00)	0.00 (0.00)

Verbal report of bias potential. Similar to what was asked of individuals in Experiment 2, teams in Experiment 4 were asked to report to the experimenter any information traps they encountered, along with the associated items or "indicators" of the trap. Table 18 shows the proportion of FAOs that were identified as decision making traps (regardless of whether the correct bias was associated with the trap). Teams were most likely to report traps in relation to the oversensitivity (50%) and availability (40%) FAOs. There were few or no traps reported for absence of evidence, randomness, and persistence of discredited information FAOs. Note that there were some FAOs reported as traps even though the participants were not specifically trained on that type of trap.

Table 18
Proportion of FAO Events Verbally Reported as a Trap (Any Bias)

FAO Event Type	% Reported as Trap (any)
Absence of Evidence	10%
Availability	40%
Oversensitivity	50%
Persistence of Discredited Information	0%
Randomness	10%
Sample Size	20%
Vividness	30%
Average	23%

Item modality. The proportion of reported data items associated with incident events is shown in Figure 11, with results separated by item modality. As was the case in Experiment 3, associated video and audio items were most likely to be included in incident reports. Likewise, image items were the least likely to be included. Although the trend is consistent with the previous experiment, a one-way repeated measures ANOVA showed no significant effect of modality in reporting items for incident events, $F(5, 20) = 0.802, p = .561$.

Due to the overall very low proportion of FAO events reported, no meaningful analysis of data item modality was possible for FAO events.

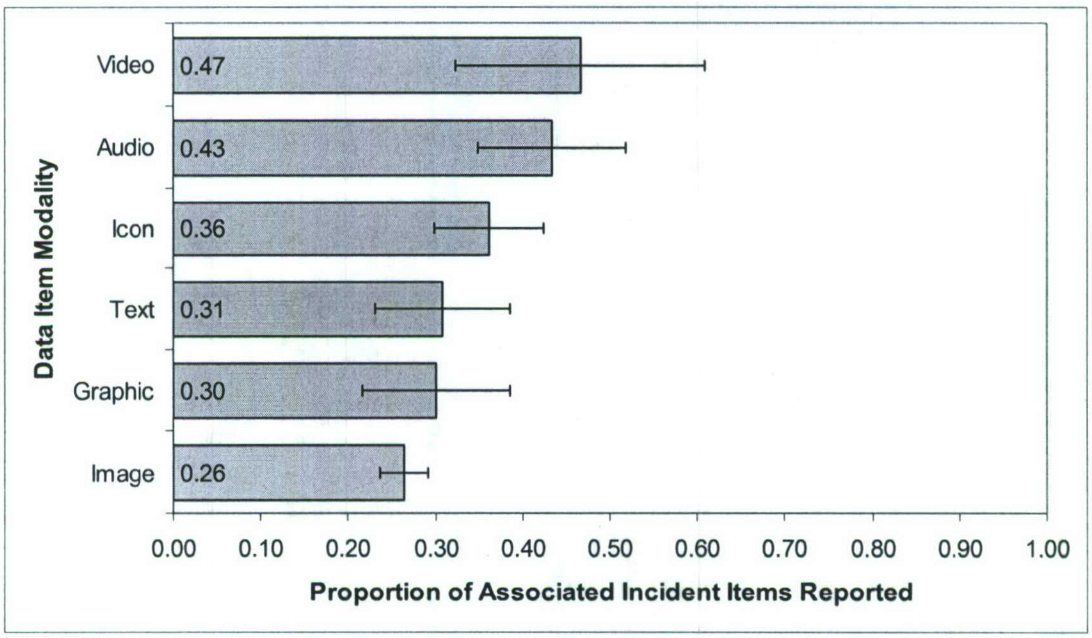

Figure 11. Proportion of Experiment 4 incident items submitted, by modality.

Response to prompts. On average, teams responded correctly to half of the prompts given (M = 0.50, SD = 0.16). Of the prompts that were answered, 74% were answered correctly, on average (SD = 0.08). As in Experiment 3, accuracy in responding to prompts seemed to vary depending on the correct answer, although in Experiment 4 the difference was not statistically significant. Prompts that were to be confirmed were answered correctly 56% of the time, and prompts that were to be denied were answered correctly 44% of the time.

Table 19 compares the mean submission rate for events that were associated with prompts versus those that were not. For each type of event, submission rates were higher when the event was associated with a prompt. Only the FAC difference was statistically significant, $F(1, 4) = 16.00$, $p = .016$. Teams were more likely to report FAC events as incidents when the events were associated with a prompt.

Table 19
Report Submission by Association With Prompt

	Mean Proportion Submitted (SD)	
Event Type	Events Associated With Prompts	Events Not Associated With Prompts
Incident	.52 (.18)	.40 (.09)
FAO	.07 (.15)	.00 (.00)
FAC	.20 (.11)	.00 (.00)

Extra attachments. About one quarter of reports filed by teams contained at least one incorrect item attachment (M = 0.23, SD = 0.08). The number of extra attachments in a report ranged from 1 to 4. The majority of extra items were items associated with other events (78%) as opposed to filler items (22%).

Of the items included in a report that were intended for other events, 50% were items of the same incident category intended for a different event (e.g., two separate sniper events reported as a single incident), and 21% were items from a very similar incident category (e.g., items for a road mine event filed with a car bomb report).

There were only four filler items submitted with reports: two "superficial similarity" filler items, one "irrelevant" filler item, and one "standard" filler item.

Experiment 4 Discussion

Experiment overview. Experiment 4 repeated the test trials and tasks performed in Experiment 3, the only manipulation being that teams performing the experiment received anti-bias training. This training was similar to that in Experiment 2, but revised for clarity.

Event type. The report scores were highest for incident events and nearly zero for FAOs and FACs. This shows that the training in Experiment 4 was effective in reducing susceptibility to making false alarms. This improvement was probably due to improvements made in training as explained in the methods section. These included the simplification of biases, task-relevant descriptions of the biases, better use of examples, improved practice trials, follow-ups on practice trials, and other factors present in the original training. In addition, the FAOs had more straightforward implementations of the bias than the FAOs used in Experiment 2. The fact that training also reduced susceptibility to FACs suggests that bias might not be the only aspect of falling for FAOs. It also suggests that training may not be specific to biases, improving the general ability for participants to recognize when something is not reportable.

Decision biases. Teams receiving anti-bias training committed very few false alarms. There was no effect of bias type on FAO percent, although the report scores were not high enough to show any difference, due to a floor effect. Still, the overall trend in FAO percent was similar to the trends observed in previous experiments, with the very small percentage of FAO events reported occurring on biases that produced higher proportions of false alarms in other experiments (oversensitivity and availability).

There was also no difference in report score between each FAO and its paired FAC, for each of the seven biases. This is probably again due to extremely low scores for both FAOs and FACs. As noted previously, it was discovered after Experiment 3 had been completed that two of the FAC events (oversensitivity and randomness) were deemed "faulty" as they could reasonably be construed as incidents. These were modified prior to Experiment 4. Of particular note is that three of the five teams reported the FAC event associated with availability as a true incident. As discussed in more detail in the Experiment 3 methods section, it was only discovered after both experiments had run that the FAC item was unintentionally converted back into a valid FAO by being associated with a prompt.

Item modality. The rank order of item modality for incident events was almost identical to the rank order observed in Experiment 3. Incident reports contained higher proportions of audio and video items, which may be more memorable and therefore more likely to be included when team members are asking for event information. Alternately, it might be the case that the most salient events or those participants had the easiest time reporting had a greater proportion of audio and video items. In this experiment it is impossible to determine the direction of influence (or to what extent the rank order is influenced by both item modality and event salience). If the rank order of item modality does indeed depend more on even salience, it might help explain why there are so many differences in rank order across experiments.

Response to prompts. In Experiment 4, prompts only seemed to impact the reporting of FAC events. Still, this effect was mostly due to the large number of false alarms associated with the availability bias described above. If construed as an FAO instead of an FAC, a significant difference is instead associated with the FAO events.

Percent reports containing extra attachments. As was observed in Experiment 3, teams largely submitted reports containing only those items that were experimentally associated with an event. Just 23% of reports submitted had one or more extra attachments. Similar to what was observed in Experiment 3, the extra items most often came from events of the same or similar categories (e.g., a graph meant for sniper event 1 was instead submitted with sniper event 2). Thus, it seems that again, extra attachments seemed to be used as confirming evidence for experimentally defined incidents or false alarms. There were no patterns suggestive of an interpretation of the data that might be the result of biases or errors outside of those under study.

Summary and Conclusions

The central objective of this research was to increase understanding of how individuals process large amounts of ambiguous data being rapidly received and to determine how to improve this ability through training. The direction of the research was guided by two major hypotheses. Our primary hypothesis was that the bias-driven characteristic errors observed in other types of decision making research would be operative in tasks involving the rapid receipt of large amounts of ambiguous data. Our second hypothesis was that targeted anti-bias training would help individuals overcome the effects of cognitive bias by teaching them to learn and recognize the markers of biased decision making. That is, training was focused on teaching individuals how to identify situations where people are inclined to poor decision making, rather than trying to teach them to eliminate the bias. The specific objectives of the research included studying the effects of overall volume of data, density of relevant data, and data format (modality) on assignment of meaning and on the types of errors and biases observed in a rapid decision making task, as well as evaluating the effectiveness of anti-bias training as a method for improving performance.

Experiment 1 was the initial investigation into the role that cognitive biases play in rapid decision making within a military context, also considering the effect of quantity and modality of information on the task. Experiment 2 extended the variables investigated in Experiment 1 to include the effect of training individuals about markers of the decision-making biases confirmed in Experiment 1. Experiment 3 considered the performance of small teams, rather than individuals, and did not administer anti-bias training. Experiment 4 was identical to Experiment 3 but added anti-bias training. This training was similar to that in Experiment 2, but revised for clarity.

Key Findings and Conclusions

Characteristic errors and decision biases. Research results supported our primary hypothesis that each of the seven types of characteristic errors of interest identified in other types of decision making research also occurs in rapid decision making. A comparison of FAO percent across the four experiments reveals some interesting trends. First, the number and types of errors made differed, depending on whether individuals performed the task independently or in teams. In Experiment 1, the FAO percent represents task performance of individuals with no training on biases (Table 4). The average error rate for individuals was as high as 50% for oversensitivity events, and as low as 26% for persistence of discredited information events. In Experiment 3, the FAO percent represents task performance of teams with no training on biases (Table 12). The average error rate for teams was as high as 70% for oversensitivity events, with no evidence of bias sensitivity on sample size events.

An examination of the rank order of FAO percent across Experiments 1-3 shows that errors associated with the oversensitivity and vividness biases were observed at higher rates in all three. In Experiment 4, where teams trained on decision making performed the task, there were almost no characteristic errors observed. Across experiments, the lowest overall rate observed was for the persistence of discredited information and sample size biases.

Participants may be particularly susceptible to the oversensitivity bias in team environments where recognizing source redundancy might require additional communication among individuals. The bias occurs when several pieces of identical information are interpreted as providing additional or corroborating evidence. It is possible that design changes in the information handling system could help guard against some instances of the oversensitivity bias. For example, a script might compare the sources of items and alert recipients to the similarities. Nonetheless, the tendency to commit errors associated with the bias seems robust in the context of rapid decision making, making it an important subject for training. The vividness bias also deserves increased attention, where users can be taught to recognize situations where the vividness of information may determine a decision more than the content of the information.

Is it possible that the characteristic errors observed were the product of something other than the decision biases identified in previous research? We tried to address this concern in the two later experiments by introducing control items for false alarm events. Each false alarm control maintained surface similarity with its FAO pair, and was modified just enough to remove the association with the underlying bias. Unfortunately, stimulus design problems made comparisons impossible for three of the seven pairs in Experiment 3, and the low overall susceptibility to errors in Experiment 4 made control comparisons irrelevant. Further research might employ similar methodology, preferably with multiple FAO/FAC pairs. Nonetheless, it is worth noting that stimulus sets contained large amounts of filler events, and that across all experiments there were very few false alarms that could not be characterized as hits on FAO events (i.e., events corresponding to the biases under study).

Items submitted with a report of an incident or FAO event that were not experimentally associated with that event were considered "extra attachments." These extra attachments were studied at length in Experiments 3 and 4. For the most part, these items showed superficial similarity to the reported event, providing further evidence that participants were for the most part sensitive only to incident and FAO events.

Training to reduce characteristic errors. A second major goal of this research program was to show that targeted anti-bias training can help to reduce the number of associated errors made in the task. Although the findings of the four experiments produced no definitive proof of training effectiveness, some interesting trends were observed. First, in Experiment 2 we observed no significant difference in FAO performance with respect to bias training, but did seem to show a general effect of "alerting"; that is, although FAO performance did not differ significantly between bias training groups, FAO performance dropped considerably in Experiment 2 as compared with Experiment 1. These results suggest that simply alerting participants to potential errors is enough to effect task performance.

We did observe an overall effect of training in the later experiments, with participants in Experiment 4 producing fewer characteristic errors than participants in Experiment 3. Still, especially given the results of Experiment 2, it is unclear whether the improvement was an effect of training, or whether similar performance could be achieved by simply alerting participants to potential errors. There was no control group included as part of Experiment 4.

In Experiment 2, participants were asked to use their anti-bias (or control) training to report potential decision making traps. The participants who received anti-bias training did report a significantly greater number of traps, but overall were not very accurate in assigning the correct bias to the trap. Although similar data were collected for Experiment 4, a technical error produced spurious data. Nonetheless, we know that traps were only rarely reported in that experiment relative to the number of opportunities.

The failure to report traps can mean several things. The obvious conclusion is that participants oftentimes had difficulty identifying traps. Alternately, the failure to report traps might be the result of waiting for all information to arrive in the inbox before conceding that the items constituted a trap. Due to the extremely fast-paced nature of the task, it is certainly plausible that many events were placed on the "back burner," perhaps leaving little time for participants to reacquaint themselves with the event and provide a verbal report before the time limit was reached. It also is worth noting here that although the experimenter is most interested in observing how participants process FAOs, the task demands force participants to concern themselves primarily with incidents. Participants were more concerned with identifying what should be reported and why than what should not be reported and why. This emphasis certainly could have impacted the chosen strategy for task completion.

Amount of information: volume and density. Comparing data across studies, it is not clear how participants responded to variations in the quantity of information delivered. In Experiment 1, increasing irrelevant information reduced the proportion of FAO responses. It was hypothesized that a reduction in accuracy was due to failures to submit reports on events that the participant was unsure about. In Experiment 2, increasing either type of information increased the likelihood of falling for a FAO. It was hypothesized that in this case, participants were more likely to just say something happened as the extraneous information increased, even if they were not sure about the status of an event, (taking a "now or never" approach). It could be that this reflects differences in bias within signal detection theory. In this case, the indecisive strategy reflects a conservative bias and the "now or never" approach reflects a liberal bias.

Experiment 2 suggested a bias shift as information levels change. In particular, participants may adopt an indecisive strategy with increasing relevant information to a certain point (similar to Experiment 1), and then adopt a "now or never" strategy when it increases even more (similar to Experiment 2). This change in strategy could be reflective of a shift from a conservative bias (if not sure, will not submit a report) to a more liberal bias (if not sure, will submit a report). Within this hypothesized framework of information volume, increasing information flow will increase either false alarms or misses, and thus one has to evaluate which is most important: minimizing false alarms or minimizing misses. Once this has been decided, participants could be trained to follow the strategy that leaders feel best suits the mission.

To prevent indecisiveness, the information management system could be designed to put warnings on report folders that have been stagnant for long periods of time. These warnings could remind the user that the time the folder has remained unaltered suggests that the user is unsure of the status of the event, and that the information needs to be carefully considered and dealt with.

Clearly there are many interpretations of the bias that participants assume in this military rapid-decision making environment, depending on how the data are considered. The best way to understand this aspect of decision making would involve directly measuring bias (beta) within the signal detection theory paradigm. This could include a manipulation of the penalty for the two types of errors (misses and false alarms), to see how that affects performance on this task. Direct measurements of bias would require more highly controlled research with many trials and direct measurements of the proportion of hits and false alarms.

Modality. As suggested in the discussion of individual experiments, it was difficult to identify any clear trends in the modality of items attached to reports across experiments. The complex experimental design used in these experiments made it difficult to balance item type across all relevant factors. There were simply too many factors involved to allow for complete counterbalancing of modality across items. For example, graphic items rarely served as key indicators simply because the information required for a key indicator was impossible to produce as a graphic. Therefore, graphic items might have lower inclusion rates in incident reports not because people ignore graphics, but because the most important item to the report was rarely a graphic item.

Some interesting trends are noted in the experiment discussions. For example, reports associated with false alarm opportunities seemed to contain relatively high proportions of items in visual data formats, such as images and icons. Still, the variability in modality of items reported was especially pronounced in FAO reports. Any reliable conclusions regarding the influence of item modality is a subject of future research.

Recommendations for Future Research

As demonstrated in previous research, the present set of experiments has demonstrated that the human decision maker is sensitive to particular conditions that promote characteristic errors, finding that the same biases identified in other types of research also occur in the context of rapid decision making brought on by the need to process large data sets in a limited time period. The research also has shown that people are sensitive to training. Participants committed fewer errors when even so much as alerted to the possibility of "ambiguities or other situations that may present an opportunity for error."

The reduced error rate is promising, but came with a cost. Training had the unintended consequence of reducing the overall hit rate on incidents. This finding suggests that when the prospect of potential errors is highlighted, people adopt a more cautious approach to the task. Although that approach produces fewer errors, it is detrimental to overall task performance. As the demand for rapid data processing in the military increases, a "cautious approach" strategy is simply unsuitable. A goal of future research, therefore, will be to devise training methods that not only reduce or eliminate characteristic errors, but do so in a way that does not detrimentally impact overall task performance (i.e., maximize the hit rate while minimizing false alarms).

There are many potential ways to revise training to accomplish this goal. The tasks and simulations devised and carried out in the C4ISR simulation laboratory for the four experiments described in this paper provide a practical approach to studying revised methods of training.

Additional types of training could be developed to not only recognize conditions that produce characteristic errors, but provide strategies for sifting through information to assist in the determination of relevance. People might benefit from training that helps them to consider not only what is relevant, but what is irrelevant.

Future studies also might attempt to minimize the "cautious approach" by providing incentives for working quickly. This could be accomplished by modifying the task to include a "moving window" of time. Participants would be required to process each item within a given amount of time before it disappeared from the inbox (i.e., they only have so much time to make a decision before it is too late for the military to address). Alternately, prompt items could ask for immediate decisions, for example "What's the final verdict for the event occurring at 3347 9932? Go or No Go?" Even the "passive" confirm/deny prompts used in Experiments 3 and 4 (which did not require a response) seemed to apply added pressure on teams to address those events that were related to prompt items.

It is possible that the efficacy of the anti-bias training in Experiments 2 and 4 could be improved simply through more practice. Although participants who received training completed a short test on the definitions of characteristic errors, that knowledge did not necessarily translate into task performance. More extensive training could include more time on task, with additional feedback on errors and missed opportunities over numerous trials.

References

Cannon-Bowers, J.A., & Salas, E. (1998). *Making Decisions under Stress: Implications For Individual and Team Training.* Washington, DC: American Psychological Association.

Fischer, D. (1970). *Historian's Fallacies*. New York: Harper Torchbooks.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). *Fault Trees: Sensitivity of Estimated Failure Probabilities to Problem Representation* (Technical Report PTR-1 042-77-8.) Eugene, OR: Decision Research.

Fleishman E.A. & Harris E.F. (1962). Patterns of leadership behavior related to employee grievances and turnover. *Personnel Psychology, 15 (2)*, 43–56.

Folds, D. J., Fain, W. B., Beers, T. M., Stocks, D. R., Coon, V. E., & Ray, J. B. (1996). *Human Factors Issues Regarding Management of Variable Message Signs in the Traffic Management Center* (Report A-9309-M.3.) Atlanta, GA: Georgia Tech Research Institute.

Folds, D. J., Fain, W. B., Stocks, D. R., Beers, T. M., & Ray, J. B. (1998). *Human Factors Issues Regarding Division of Labor in the Traffic Management Center* (Report A-9309-M.4). Atlanta, GA: Georgia Tech Research Institute.

Folds, D. J., Mitta, D. A., Fain, W. B., Beers, T. M., & Stocks, D. R. (1995*). Human Factors Issues Regarding Incident Detection Support Systems in the Traffic Management Center* (Report A-9309-M.2.) Atlanta, GA: Georgia Tech Research Institute.

Gettys, C., Kelly, C. W, III, & Peterson, C. (1973). The best guess hypothesis in multistage inference. *Organizational Behavior and Human Performance, 10,* 363-373.

Heuer, Richards J., Jr. (1999). *Psychology of Intelligence Analysis*. Washington, DC: Center for the Study of Intelligence, Central Intelligence Agency.

Kahneman, D., & Tversky, A. (1972). Subjective probability: a judgment of representativeness. *Cognitive Psychology*, *3*, 430-454.

Klein, G.A. (1989). Recognition-primed decisions. In W.B. Rouse (Ed.), *Advances in Man-Machine Systems Research, Vol. 5.* Greenwich, CT: JAI Press, 47-92.

Lau, R., Lepper, M., & Ross, L. (1976). Resistance of inaccurate and discredited personal impressions: a field demonstration of attributional perseverance. Paper presented at 56th Annual Meeting of the Western Psychological Association.

Nisbett, R. & Ross, L. (1980). *Human Inference: Strategies and Shortcomings of Social Judgment.* Englewood Cliffs, NJ: Prentice-Hall.

Simon, H. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics, 69*, 99-118.

Tversky, A., & Kahneman, D., (1973). Availability: A heuristic for judging frequency and probability, *Cognitive Psychology, 5*, 207-232.

Tversky, A., & Kahneman, D., (1974). Judgment under uncertainty: heuristics and biases. *Science, 185*, 1124-1131.